Supplemental Material

Supplemental note

Derivation of odds ratio transformations OR_1 and OR_2

Letting p_0 and p_1 represent the frequency of the risk allele (or effect allele) within controls and cases respectively we can write the OR as

$$OR = \frac{p_1}{1-p_1}\frac{1-p_0}{p_0}. \tag{1}$$

If we have individual-level data, then we can estimate p_0 and p_1 from the sample and calculate the OR directly using (1), without making any further assumptions. However, if only summary statistics are available we seek to derive an expression for OR that potentially depends on summary statistics generated from a linear regression model.

We assume the following simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \tag{2}$$

where Y_i is the response variable for individual $i = 1, \ldots, n$ of a population, which we assume takes values 0 or 1 for unaffected (controls) and diseased (cases) individuals respectively. We define K as the lifetime probability that an individual will be affected by the disease in the population. By definition $\mathbb{E}(Y_i) = \mathbb{P}(Y_i = 1) = K$, where the $\mathbb{E}$ notation denotes expectation. The independent predictor variable X_i is considered random and models a SNP. The random variable X_i takes values 0, 1, or 2 with the corresponding allele frequency of the risk allele, denoted p, and we assume that each SNP is independent. The random variable X_i is thus Binomial$(2, p)$ distributed for each SNP. In Equation (2), ϵ_i is a random error term such that $\mathbb{E}(\epsilon_i) = 0$ and $\text{Var}(\epsilon_i|X_i) = \sigma^2$ and the unknown parameters β_0 and β_1 are to be estimated.

Under the simple linear regression model (2), we seek to solve for expressions of p_0 and p_1 and substitute these into (1) to complete our derivation of the transformation. We use the expression for the ordinary least squares estimator of β_1, and $p = (1-k)p_0 + kp_1$ (from the law of total probability) to back solve for p_0 and p_1 using the following relationships:

$$\beta_1 = \frac{\text{Cov}(Y, X)}{\text{Var}(X)} \tag{3}$$

$$\text{Cov}(Y, X) = 2k(1-k)(p_1 - p_0) \tag{4}$$

$$\text{Var}(X) = \text{Var}(X|Y=0)(1-k) + \text{Var}(X|Y=1)k + 4k(1-k)(p_0 - p_1)^2, \tag{5}$$

with $\text{Var}(X|Y=0)$ and $\text{Var}(X|Y=1)$ equaling the variances of the coded genotypes in the controls and cases respectively. To derive the covariance we have

$$\begin{aligned}\text{Cov}(Y,X) &= \mathbb{E}(YX) - \mathbb{E}(Y)\mathbb{E}(X)\\ &= \mathbb{E}(YX) - 2kp.\end{aligned}$$

As we assume that Y takes values 0 and 1 and X takes values 0, 1, and 2 the only product outcomes that contribute to $\mathbb{E}(YX)$ are when $Y = 1$ and $X = 1$, and when $Y = 1$ and $X = 2$. Therefore,

$$\begin{aligned}\mathbb{E}(XY) &= \sum_{x=0}^{2}\sum_{y=0}^{1} xy\mathbb{P}(X = x, Y = y)\\ &= \mathbb{P}(X = 1, Y = 1) + 2\mathbb{P}(X = 2, Y = 1)\\ &= \mathbb{P}(X = 1|Y = 1)\mathbb{P}(Y = 1) + 2\mathbb{P}(X = 2|Y = 1)\mathbb{P}(Y = 1)\\ &= 2p_1(1 - p_1)k + 2p_1^2k\\ &= 2p_1k, \text{ therefore}\\ \text{Cov}(Y,X) &= 2p_1k - 2kp. \qquad (6)\end{aligned}$$

It is noted that the above derivation of the $\mathbb{E}(XY)$ and subsequent $\text{Cov}(Y,X)$ required the assumption that the within case genotype frequencies are in Hardy-Weinberg equilibrium (HWE). Substituting $p = (1 - k)p_0 + kp_1$ into (6) we have

$$\begin{aligned}\text{Cov}(Y,X) &= 2p_1k - 2k[(1 - k)p_0 + kp_1]\\ &= 2p_1k - 2kp_0 + 2k^2p_0 - 2k^2p_1\\ &= 2k(1 - k)(p_1 - p_0). \qquad (7)\end{aligned}$$

To derive equation (5), we note that the events $Y = 1$ and $Y = 0$ partition the whole outcome space and thus we can use the law of total variance to obtain

$$
\begin{aligned}
\mathrm{Var}(X) = \sum_{y=0}^{1} \mathrm{Var}(X|Y=y)\mathbb{P}(Y=y) + \sum_{y=0}^{1} \mathbb{E}(X|Y=y)^2[1-\mathbb{P}(Y=y)]\mathbb{P}(Y=y) - \\
2\sum_{y=1}^{1}\sum_{j=0}^{1-1} \mathbb{E}(X|Y=y)\mathbb{P}(Y=y)\mathbb{E}(X|Y=j)\mathbb{P}(Y=j) \\
&= \mathrm{Var}(X|Y=0)\mathbb{P}(Y=0) + \mathrm{Var}(X|Y=1)\mathbb{P}(Y=1) + \\
&\quad \mathbb{E}(X|Y=0)^2[1-\mathbb{P}(Y=0)]\mathbb{P}(Y=0) + \\
&\quad \mathbb{E}(X|Y=1)^2[1-\mathbb{P}(Y=1)]\mathbb{P}(Y=1) - \\
&\quad 2\mathbb{E}(X|Y=1)\mathbb{P}(Y=1)\mathbb{E}(X|Y=0)\mathbb{P}(Y=0) \\
&= \mathrm{Var}(X|Y=0)(1-k) + \mathrm{Var}(X|Y=1)k \\
&\quad + 4p_0^2 k(1-k) + 4p_1^2(1-k)k - 8p_0p_1(1-k)k \\
&= \mathrm{Var}(X|Y=0)(1-k) + \mathrm{Var}(X|Y=1)k + 4k(1-k)(p_0-p_1)^2. \qquad (8)
\end{aligned}
$$

We cannot observe the $\mathrm{Var}(X)$ or $\mathrm{Var}(X|Y=0)$ and $\mathrm{Var}(X|Y=1)$ from summary statistics and thus we must make some assumptions about the form of the variance for the SNP. Initially, we can assume that the SNP genotype frequencies across cases and controls are in HWE and let $\mathrm{Var}(X) = 2p(1-p)$. This assumes that $2p(1-p)$ does not deviate substantially from (8). In the following section (*Deviation between two variance assumptions*) we show when this assumption is reasonable. If we make this assumption, then we can combine equations (3), (4) and $\mathrm{Var}(X) = 2p(1-p)$ to arrive at

$$\beta_1 = \frac{2k(1-k)(p_1-p_0)}{2p(1-p)},$$

which implies

$$p_1 = p_0 + \frac{\beta_1 p(1-p)}{k(1-k)}. \qquad (9)$$

Substituting,

$$p_0 = \frac{p-kp_1}{(1-k)} \qquad (10)$$

from $p = (1-k)p_0 + kp_1$ into (9) then

$$\begin{aligned} p_1 &= \frac{p - kp_1}{1-k} + \frac{\beta_1 p(1-p)}{k(1-k)} \\ p_1 &= p + \frac{\beta_1 p(1-p)}{k}. \end{aligned} \tag{11}$$

Combining equation (1), (10) and (11) we have the following expression for the transformation to OR

$$\begin{aligned} OR_1 &= \frac{\left[\frac{pk+\beta_1 p(1-p)}{k}\right]\left[1 - \frac{p-pk-\beta_1 p(1-p)}{1-k}\right]}{\left[1 - \frac{pk+\beta_1 p(1-p)}{k}\right]\left[\frac{p-pk-\beta_1 p(1-p)}{1-k}\right]} \\ &= \frac{[k+\beta_1(1-p)][1-k+\beta_1 p]}{[k-\beta_1 p][1-k-\beta_1(1-p)]}, \end{aligned} \tag{12}$$

which corresponds to (5) in the main text.

We can improve on the assumption of $\mathrm{Var}(X) = 2p(1-p)$ by assuming the genotype frequencies within cases and controls are in HWE. Under this assumption, we let $\mathrm{Var}(X|Y=0) = 2p_0(1-p_0)$ and $\mathrm{Var}(X|Y=1) = 2p_1(1-p_1)$ in (8) and substitute the variance into (3) to obtain

$$\beta_1 = \frac{k(1-k)(p_1-p_0)}{[p_0(1-p_0)(1-k) + p_1(1-p_1)k + 2k(1-k)(p_0-p_1)^2]}. \tag{13}$$

Letting $A = k(1-k)$ and $B = (1-k)$ in equation (13) we have

$$\begin{aligned} Ap_1 - Ap_0 =& \beta_1 B[p_0 - p_0^2] + \beta_1 k[p_1 - p_1^2] + 2A\beta_1[p_0^2 - 2p_0p_1 + p_1^2] \\ 0 =& p_0^2[2A\beta_1 - \beta_1 B] + p_0[A + \beta_1 B] - 4A\beta_1 p_0 p_1 + p_1^2[2A\beta_1 - \beta_1 k] + p_1[-A + \beta_1 k]. \end{aligned}$$

We know $p = (1-k)p_0 + kp_1$ and thus substituting $[p - Bp_0]/k$ for p_1

$$\begin{aligned} 0 =& p_0^2\left[2A\beta_1 - \beta_1 B + \frac{4AB\beta_1}{k} + \frac{B^2[2A\beta_1 - \beta_1 k]}{k^2}\right] + \\ & p_0\left[A + \beta_1 B - \frac{4A\beta_1 p}{k} - \frac{2Bp[2A\beta_1 - \beta_1 k]}{k^2} - \frac{B[-A+\beta_1 k]}{k}\right] + \\ & \frac{p^2[2A\beta_1 - \beta_1 k]}{k^2} + \frac{p[-A+\beta_1 k]}{k}. \end{aligned} \tag{14}$$

This can be simplified to a quadratic in p_0 with the following coefficients

$$a = \frac{\beta_1(1-k)}{k}, \tag{15}$$

$$b = \frac{(1-k)(k-2\beta_1 p)}{k}, \tag{16}$$

$$c = \frac{p^2\beta_1 - 2p^2\beta_1 k + pk^2 - pk + pk\beta_1}{k}. \tag{17}$$

Substituting these coefficients into $\frac{-b\pm\sqrt{b^2-4ac}}{2a}$ and simplifying we have the following solutions for p_0

$$p_0 = \frac{(1-k)(2\beta_1 p - k) \pm \sqrt{(1-k)[k^2(1-k)+4p\beta_1^2 k(p-1)]}}{2\beta_1(1-k)}. \tag{18}$$

Substituting $p_1 = [p-(1-k)p_0]/k$ into equation (1) and simplifying we have

$$OR = 1 + \frac{p-p_0}{p_0(k-p)+p_0^2(1-k)}. \tag{19}$$

Substituting (18) into equation (19) we have the following transformation (after some algebra)

$$OR_2 = 1 + \frac{2\beta_1\left\{2p\beta_1(1-k)-(1-k)(2\beta_1 p-k)\pm\sqrt{k(1-k)[k(1-k)-4p(1-p)\beta_1^2]}\right\}}{2\beta_1(k-p)\left\{(1-k)(2\beta_1 p-k)\pm\sqrt{k(1-k)[k(1-k)-4p(1-p)\beta_1^2]}\right\}+\left\{(1-k)(2\beta_1 p-k)\pm\sqrt{k(1-k)[k(1-k)-4p(1-p)\beta_1^2]}\right\}^2}.$$

Deviation between two variance assumptions

We seek to characterise the deviation between the two variance assumptions, which is the primary difference between the derivation of OR_1 and OR_2. We define

$$\begin{aligned}
\mathrm{Var}(X)_1 &= 2p(1-p) = 2[(1-k)p_0+kp_1][1-(1-k)p_0-kp_1] \\
&= 2p_0 - 2p_0^2 + 2kp_0^2 - 2kp_0p_1 - 2kp_0 + 2kp_0^2 - 2k^2p_0^2 + 2k^2p_0p_1 \\
&+ 2kp_1 - 2kp_1p_0 + 2k^2p_1p_0 - 2k^2p_1^2, \\
\mathrm{Var}(X)_2 &= 2p_0(1-p_0)(1-k) + 2p_1(1-p_1)k + 4p_0^2k(1-k) + 4p_1^2(1-k)k - 8p_0p_1(1-k)k \\
&= 2p_0 - 2p_0^2 - 2kp_0 + 2kp_0^2 + 2kp_1 - 2kp_1^2 + 4p_0^2k - 4p_0^2k^2 + 4kp_1^2 - 4p_1^2k^2 \\
&- 8kp_0p_1 + 8p_0p_1k^2.
\end{aligned}$$

Looking at the difference and cancelling like terms

$$\mathrm{Var}(X)_1 - \mathrm{Var}(X)_2 = 2p_0^2(k^2-k) + 2p_1^2(k^2-k) + 4p_0p_1(k-k^2).$$

Let $d(p_0, p_1, k) = 2p_0^2(k^2-k) + 2p_1^2(k^2-k) + 4p_0p_1(k-k^2)$ be the difference function. This equation will always be negative, which can be shown by factorising this function such that

$$d(p_0, p_1, k) = 2(k^2-k)(p_0-p_1)^2.$$

Because $0 \leq k \leq 1$ then k^2 will always be less than k, which implies that we have a negative term multiplied by a squared term, which is always negative. This implies that $\mathrm{Var}(X)_2$ is always greater than $\mathrm{Var}(X)_1$. We wish to characterise the maxima and minima of this function subject to the constraints that $0 \leq p_0 \leq 1, 0 \leq p_1 \leq 1, 0 \leq k \leq 0.5$. This will allow us to characterise when these variance assumptions deviate substantially. Equating the partial derivatives of $d(p_0, p_1, k)$ to 0 we have

$$\begin{aligned}
d'_{p_0}(p_0, p_1, k) &= 4p_0(k^2-k) - 4p_1(k^2-k) = 0,\\
d'_{p_1}(p_0, p_1, k) &= 4p_1(k^2-k) - 4p_0(k^2-k) = 0,\\
d'_k(p_0, p_1, k) &= 4p_0^2k - 2p_0^2 + 4p_1^2k - 2p_1^2 + 4p_0p_1 - 8p_0p_1k = 4k(p_0-p_1)^2 - 2(p_0-p_1)^2 = 0,
\end{aligned}$$

which leads to

$$\begin{aligned}
&4p_0(k^2-k) - 4p_1(k^2-k) = 0 \Rightarrow p_0 = p_1,\\
&4p_1(k^2-k) - 4p_0(k^2-k) = 0 \Rightarrow p_1 = p_0,\\
&4k(p_0-p_1)^2 - 2(p_0-p_1)^2 = 0 \Rightarrow k = 1/2.
\end{aligned}$$

When $p_0 = p_1$ there is no effect and the variance difference is maximised at 0. For $k = 1/2$ we check the boundary to locate the minimum. Substituting $k = 0.5$ into $d(p_0, p_1, k)$ to obtain

$$d(p_0, p_1, 1/2) = -\frac{1}{2}(p_0-p_1)^2,$$

which is minimised when the difference between p_0 and p_1 is greatest or when $(p_0, p_1) = (0, 1)$ and $(p_0, p_1) = (1, 0)$. These points are both global minima and imply that the greatest difference between OR_1 and OR_2 is achieved when $k = 0.5$ and the difference in case-control allele frequency (assuming genotypes are in HWE) for the risk allele is at its greatest. This implies that under the assumed model that the transformation is derived under, we expect OR_1 and OR_2 to be most different when k approaches 0.5 and when the effect is large. Transformation OR_2 should perform equally well or better than OR_1 if the data are generated under the assumed model.

Existence of root of p_0 quadratic in $(0,1)$. As stated in the main text, the solution to the quadratic in p_0 in (18) can have two, one, or no solution. By the intermediate value theorem (IVT) the quadratic polynomial

$$f(p_0) = p_0^2\frac{\beta_1(1-k)}{k} + p_0\frac{(1-k)(k-2\beta_1 p)}{k} + \frac{p^2\beta_1 - 2p^2\beta_1 k + pk^2 - pk + pk\beta_1}{k},$$

has a root in $(0,1)$ if $f(0) < 0 < f(1)$. We note that because $f(p_0)$ is a quadratic then this is a condition for a single root. Therefore, we have the following conditions

$$\frac{p^2\beta_1 - 2p^2\beta_1 k + pk^2 - pk + pk\beta_1}{k} < 0$$

which implies that

$$\beta_1 < \frac{pk - pk^2}{p^2 - 2p^2k + pk},$$

and

$$\frac{\beta_1(1-k)}{k} + \frac{(1-k)(k-2\beta_1 p)}{k} + \frac{p^2\beta_1 - 2p^2\beta_1 k + pk^2 - pk + pk\beta_1}{k} > 0.$$

Simplifying this we have

$$\begin{aligned} \beta_1(1-k) + (1-k)(k-2\beta_1 p) &> -p^2\beta_1 + 2p^2\beta_1 k - pk^2 + pk - pk\beta_1 \\ \beta_1 &> \frac{pk - pk^2 - k(1-k)}{(1-k) - 2(1-k)p + p^2 - 2p^2k + pk]}. \end{aligned}$$

The combined conditions result in the following bounds on β_1

$$\frac{pk - pk^2 - k(1-k)}{[(1-k) - 2(1-k)p + p^2 - 2p^2k + pk]} < \beta_1 < \frac{pk - pk^2}{p^2 - 2p^2k + pk}.$$

It is possible that there exists two roots to the system for which the above bounds do not hold. We will seek to find when there is no solution to the quadratic and then two solutions will be the complement of the above bounds and the no solution constraints. There will exist no solution if $b^2 - 4ac < 0$. Substituting the coefficients (15), (16), and (17) we have

$$\left[\frac{(1-k)(k-2\beta_1 p)}{k}\right]^2 - 4\frac{\beta_1(1-k)}{k}\frac{p^2\beta_1 - 2p^2\beta_1 k + pk^2 - pk + pk\beta_1}{k} < 0,$$

which simplifies to the constraints

$$-\sqrt{\frac{-k^2(1-k)}{4p^2k-4pk}} < \beta_1 < \sqrt{\frac{-k^2(1-k)}{4p^2k-4pk}}. \tag{20}$$

The numerator and denominator for (20) will always be negative (overall positive) because $k > 0$ and $4p^2k < 4pk$ because $p \in (0, 1)$.

Eliminating p

It may be the case that the allele frequency for each SNP is not reported and an adequate reference data set is not obtainable, for example, in an admixed population. If this is the case, then we can use the information contained in the standard error of the regression coefficient $se(\hat{\beta}_1)$, which is often reported with summary statistics, to derive expressions that are independent of p with equivalent assumptions to OR_1 and OR_2. We explore their adequacy relative to the expressions that include p through simulation.

The standard error of the estimator $\hat{\beta}_1$ from ordinary least squares can be represented as

$$\begin{aligned} se(\hat{\beta}_1)^2 &= \frac{\mathrm{Var}(Y) + \beta_1^2\mathrm{Var}(X) - 2\beta_1\mathrm{Cov}(X,Y)}{(n-2)\mathrm{Var}(X)} \\ &= \frac{\mathrm{Var}(Y) - \beta_1^2\mathrm{Var}(X)}{(n-2)\mathrm{Var}(X)}. \end{aligned} \tag{21}$$

From summary statistics, we usually have an estimate of the standard error of the genetic effect for each SNP and thus we can couple the expression for the ordinary least squares regression coefficient with the expression for the standard error to back solve for p_0 and p_1. Therefore, we know from (3),

$$\mathrm{Var}(X) = \frac{2k(1-k)(p_1-p_0)}{\beta_1}\ . \tag{22}$$

Rearranging (21) for $\mathrm{Var}(X)$ and using $\mathrm{Var}(Y) = k(1-k)$

$$\mathrm{Var}(X) = \frac{k(1-k)}{se(\hat{\beta}_1)^2(n-2)+\beta_1^2}. \tag{23}$$

Equating (22) and (23) we solve for $p_1 - p_0$

$$p_1 - p_0 = \frac{\beta_1}{2[se(\hat{\beta}_1)^2(n-2)+\beta_1^2]}. \tag{24}$$

The right hand side of (24) is a function of observable quantities but we require another relationship to separate p_0 from p_1. Let $d = p_1 - p_0$, then if we assume that the effect is small then $\mathrm{Var}(X) = 2p(1-p)$ and we can use $p = (1-k)p_0 + kp_1$ to show that

$$2p(1-p) = 2(p_0 + kd - p_0^2 - 2kp_0d - k^2d^2). \tag{25}$$

Equating (25) to (22) we have the following polynomial in p_0

$$-\beta_1 p_0^2 + p_0(\beta_1 - 2\beta_1 kd) + \beta_1 kd(1-kd) - k(1-k)d = 0. \tag{26}$$

Equation (26) can be solved with the quadratic formula and the following coefficients

$$\begin{aligned} a &= -\beta_1, \\ b &= \beta_1 - 2\beta_1 kd, \\ c &= \beta_1 kd(1-kd) - k(1-k)d, \end{aligned} \tag{27}$$

and then p_1 can be solved using equation (24) and the odds ratio estimated using (1).

In the same manner as above, we solve for p_0 assuming equation (8) for the variance, which does not assume HWE across cases and controls. Let,

$$\begin{aligned} \beta_1 = & \frac{2k(1-k)d}{2p_0(1-p_0)(1-k) + 2p_1(1-p_1)k + 4k(1-k)d^2} \\ & \beta_1[p_0(1-p_0)(1-k) + p_1(1-p_1)k + 2k(1-k)d^2] = k(1-k)d \\ & \beta_1 p_0(1-p_0)(1-k) + \beta_1 p_1(1-p_1)k + \beta_1 2k(1-k)d^2 - k(1-k)d = 0 \\ & -p_0^2\beta_1 + (\beta_1 - 2dk\beta_1)p_0 + \beta_1 2k(1-k)d^2 - k(1-k)d + k\beta_1(d-d^2) = 0. \end{aligned} \tag{28}$$

Equation (28) can be solved for p_0 using the quadratic formula and the following coefficients

$$a = -\beta_1, \tag{29}$$

$$b = \beta_1 - 2\beta_1 kd, \tag{30}$$

$$c = 2\beta_1 k(1-k)d^2 - k(1-k)d + k\beta_1(d-d^2), \tag{31}$$

and then p_1 can be solved using equation (24) and the odds ratio estimated using equation (1). We note that the difference in the quadratics (26) and (28) only changed in c. If we take the difference (31) $-$ (27) then we arrive at $-k\beta_1(d-d^2)$.

There will be no solution to (28) when $b^2 - 4ac < 0$. Using (29), (30) and (31) then

$b^2 - 4ac < 0$ implies that

$$\beta_1 - 4\beta_1 k^2 d^2 + 4\beta_1 d^2 k - 4kd + 4k^2 d < 0.$$

Substituting d and simplify to acquire a quadratic in the variance

$$\begin{aligned}&[se(\hat{\beta}_1)^2(n-2) + \beta_1^2]^2 - \beta_1^2 k^2 + \beta_1^2 k - 2k[se(\hat{\beta}_1)^2(n-2) + \beta_1^2] + 2k^2[se(\hat{\beta}_1)^2(n-2) + \beta_1^2] < 0\\ &se(\hat{\beta}_1)^4(n-2)^2 + se(\hat{\beta}_1)^2[2\beta_1^2(n-2) - 2k(n-2) + 2k^2(n-2)] + \beta_1^4 + \beta_1^2 k^2 - \beta_1^2 k < 0.\end{aligned} \tag{32}$$

For $n > 2$, (32) is concave down due to the positivity of the first coefficient. This implies that (32) will only be negative if the quadratic evaluated at the critical value is < 0. To find the critical value we calculate the first oder derivative of (32) to arrive at

$$se(\hat{\beta}_1)^2 = \frac{k - k^2 - \beta_1^2}{n-2}. \tag{33}$$

If (32) evaluated at (33) is less than zero and $se(\hat{\beta}_1)^2$ is within the bounds set by the following solutions to the roots of (32)

$$\frac{-[\beta_1^2 - k(1-k)] \pm \sqrt{k^2(1-k)^2 - \beta_1^2 k(1-k)}}{(n-2)},$$

then the standard error transformation breaks down.

Sampling variance of transformed regression coefficient under OR_1

For the approximate sampling variance of the estimate of β_1 on the log odds ratio scale, defined to be β_{1l}, we can use a Taylor series (ignoring third moments and greater) expansion around the true β_1 (as stated in Lynch *et al.* (1998))

$$\mathrm{Var}(\hat{\beta}_{1l}) \approx \left(\frac{\partial \beta_{1l}}{\partial \beta_1}\right)^2\Bigg|_{\hat{\beta}_1} \mathrm{Var}(\hat{\beta}_1) \tag{34}$$

where we assume $\beta_{1l} = \log(OR_1)$. We take the first derivative of the natural logarithm of equation (12)

$$\frac{\partial \log(OR_1)}{\partial \beta_1} = \frac{p(1-p)}{pk + \beta_1 p(1-p)} + \frac{p(1-p)}{1-k-p+pk+\beta_1 p(1-p)} + \tag{35}$$

$$\frac{p(1-p)}{k - pk - \beta_1 p} + \frac{p(1-p)}{p - pk - \beta_1 p(1-p)} \tag{36}$$

$$= \frac{(1-p)}{k + \beta_1(1-p)} + \frac{p}{1-k+\beta_1 p} + \frac{p}{k - \beta_1 p} + \frac{(1-p)}{1-k-\beta_1(1-p)} \tag{37}$$

Substituting this into equation (34) we can make approximate inference about the sampling variance of the estimated genetic effect on the log odds ratio scale.

Literature Cited

Lynch, M., B. Walsh, *et al.*, 1998 *Genetics and analysis of quantitative traits*, volume 1. Sinauer Sunderland, Massachusetts.

Pirinen, M., P. Donnelly, and C. C. Spencer, 2013 Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. The Annals of Applied Statistics pp. 369–390.

Supplemental figures

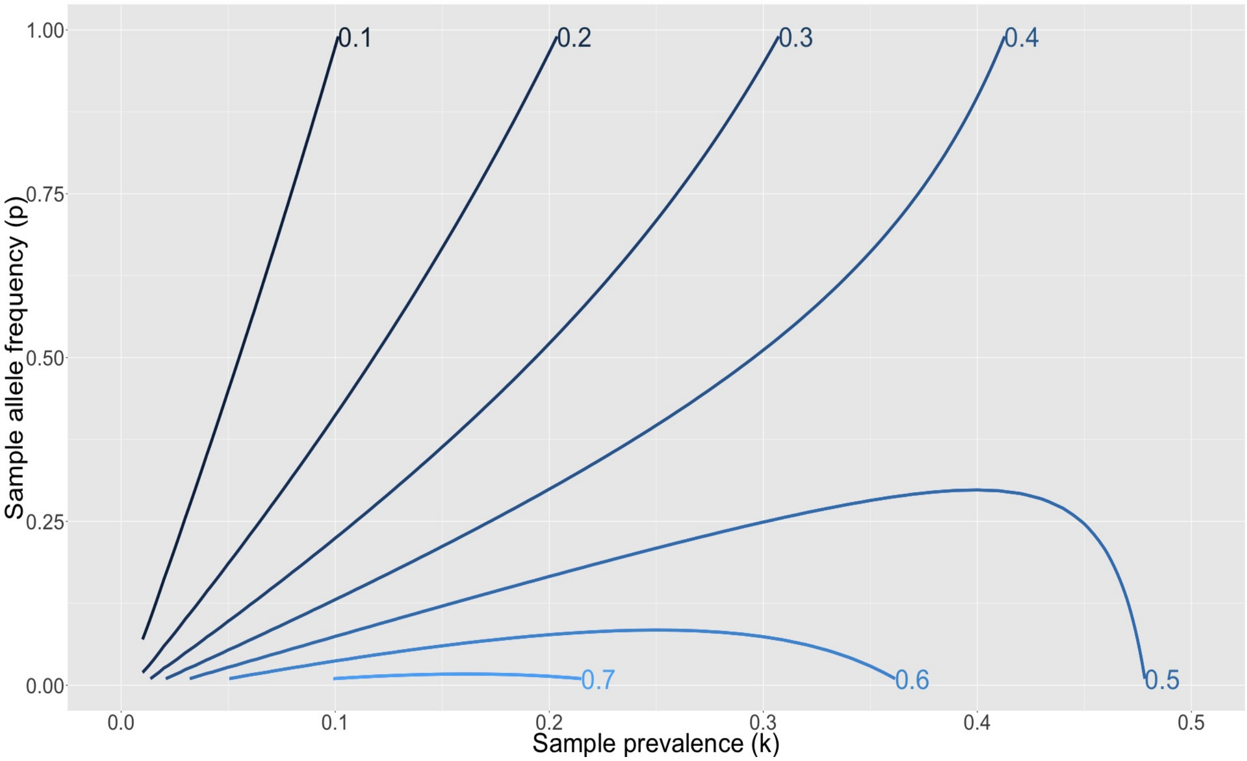

Figure S1 Contour boundaries for the LMM effect size that corresponds to an odds ratio equal to 50 for OR_2. For a particular k and p the LMM effect size that corresponds to an odds ratio equal to 50 will vary. For example, for a sample prevalence of 0.5 for all values of p the LMM effect that corresponds to an odds ratio of 50 will lie between 0.4 and 0.5. For a given k and p all effect sizes displayed lie within the bounds for the existence of one root to (18).

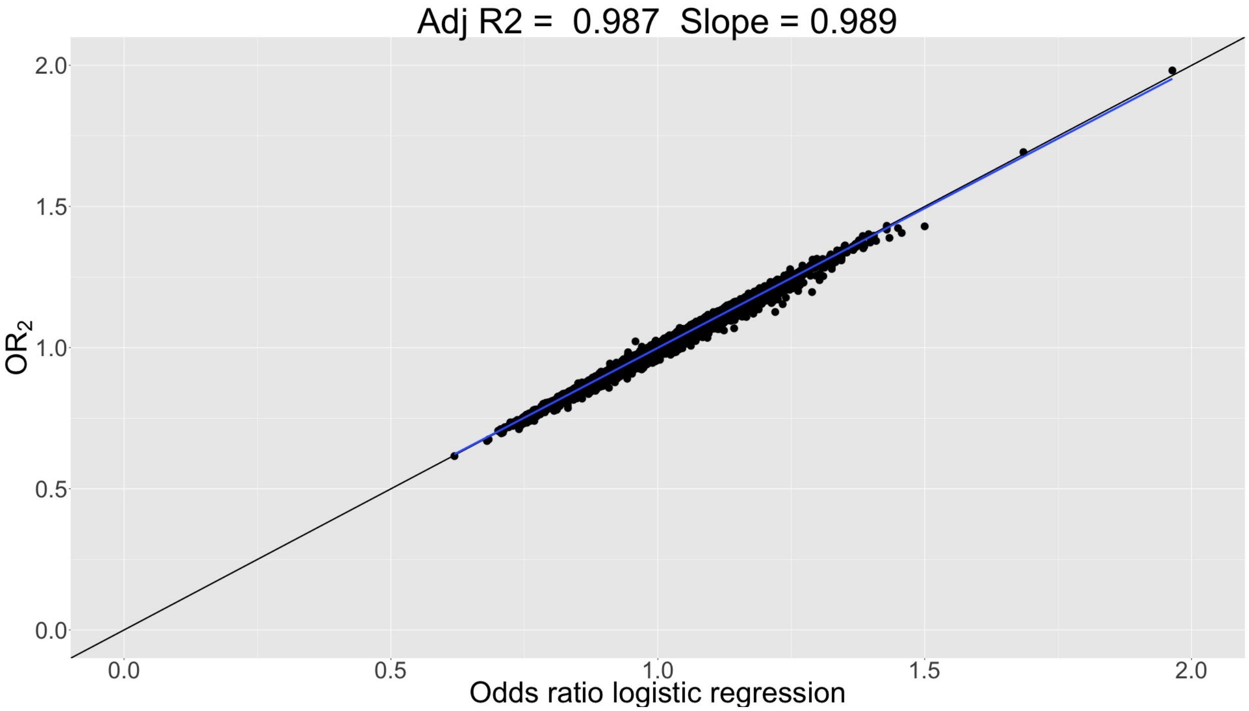

Figure S2 Performance of odds ratio transformation for null phenotype simulation. Comparison of OR_2 with estimated odds ratios from logistic regression. Phenotypes were simulated with only the effect of PC 1 and no genetics effects. Panels display comparisons of 10,000 randomly sampled loci from each of the replicates. Figure includes the fitted regression line (blue) and $y = x$ line (black) for reference with the key statistics of this regression displayed at the top of each panel.

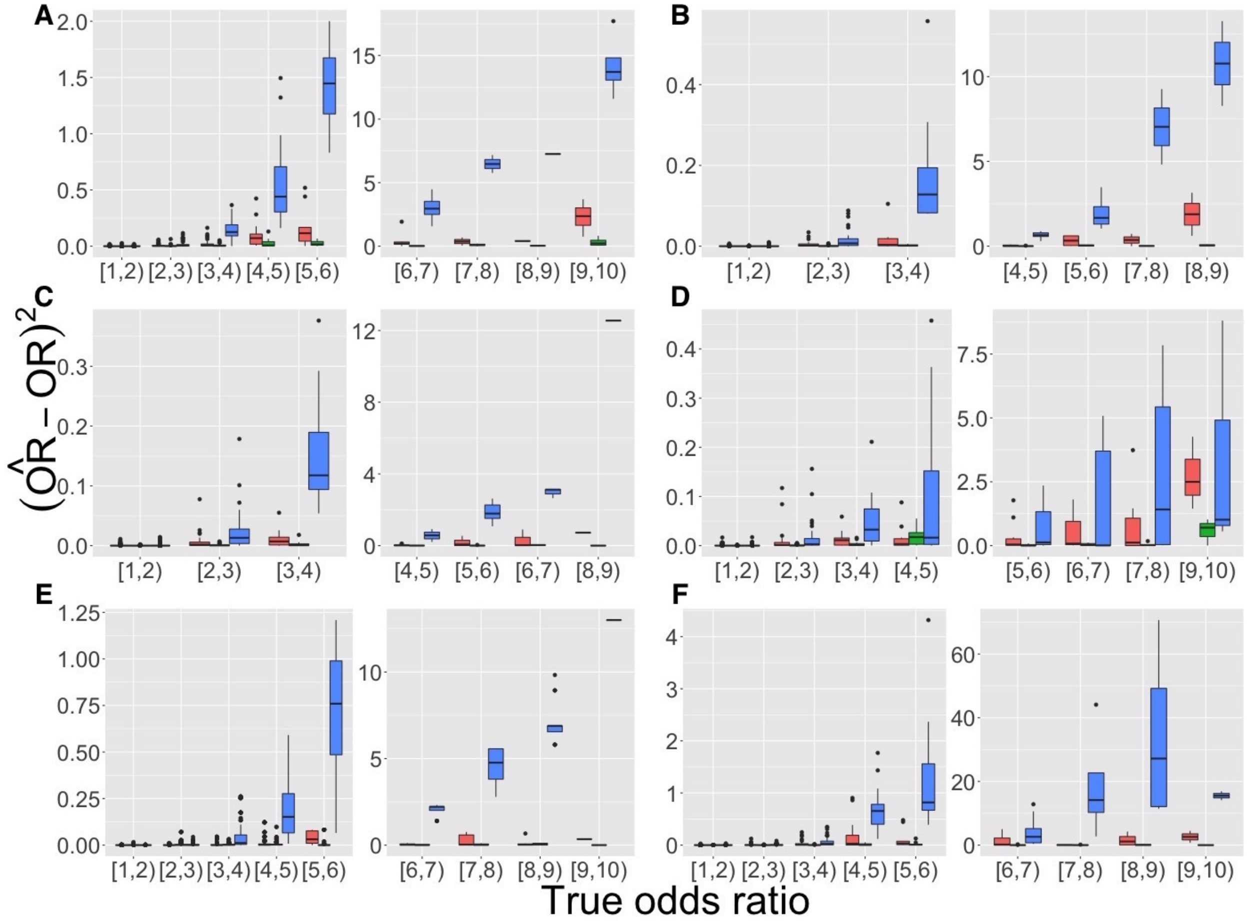

Figure S3 Performance of logistic regression and odds ratio transformations from the linear model across simulation scenarios evaluated using the squared difference from the true estimate. Comparison of squared deviations of estimated odds from the true simulated odds ratios from logistic regression (green), transformed odds ratios from the LMM using OR_2 (red), and the transformed odds ratios from the LMM using the equation from Pirinen *et al.* (2013) (blue) across logistic and liability threshold model simulation scenarios. This plot should be interpreted in conjunction with Table S1, which summarises the number of variants contributing to each bin and average allele frequency as they are not constant across bins. For each simulation scenario odds ratios were grouped into one unit bins that included the lower bin value but not the upper bin value. Each panel is split into two facets to allow for clearer comparison of the squared deviations for smaller effect sizes. Panel (A) depicts results from the logistic model simulation. Panel (B) shows results from the simulation scenario with $K = 0.1$, $h^2 = 0.5$, $n_{controls} = 5000$, and $n_{cases} = 5000$ ($k = 0.5$). Panel (C) shows results for the simulation scenario with $K = 0.05$, $h^2 = 0.5$, $n_{controls} = 5000$, and $n_{cases} = 5000$ ($k = 0.5$). Panel (D) presents results for the simulation scenario with $K = 0.02$, $h^2 = 0.5$, $n_{controls} = 8000$, and $n_{cases} = 2000$ ($k = 0.2$). Panel (E) portrays results for the simulation scenario with $K = 0.01$, $h^2 = 0.8$, $n_{controls} = 9000$, and $n_{cases} = 1000$ ($k = 0.1$). Panel (F) depicts results from the rare variant simulation scenario with $K = 0.01$, $h^2 = 0.05$, $n_{controls} = 8600$, and $n_{cases} = 1400$ ($k = 0.14$).

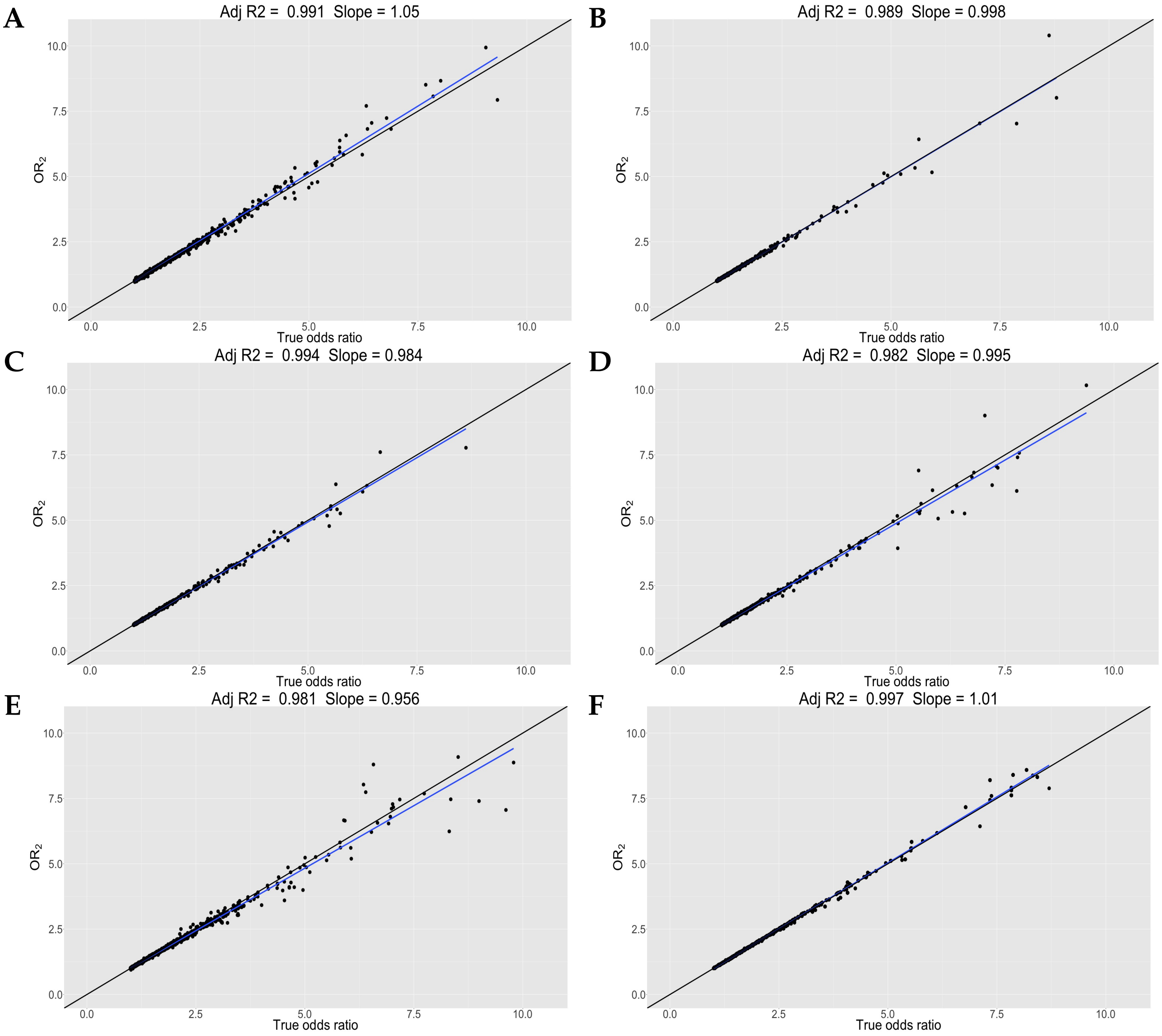

Figure S4 **Performance of odds ratio transformation when using reference allele frequencies.** Comparison of estimated odds ratios using OR_2 and reference allele frequencies from the 1000 Genomes Phase 1 Version 3 European sample with true simulated odds ratios from the logistic model simulation and liability threshold model simulation. Panel (A) depicts estimated odds ratios from the LMM on the true simulated odds ratios for the logistic model simulation. Panel (B) depicts estimated odds ratios from the LMM on the true simulated odds ratios for the simulation scenario with $K = 0.1$, $h^2 = 0.5$, $n_{controls} = 5000$, and $n_{cases} = 5000$ ($k = 0.5$). Panel (C) shows estimated odds ratios from the LMM on the true simulated odds ratios for the simulation scenario with $K = 0.5$, $h^2 = 0.5$, $n_{controls} = 5000$, and $n_{cases} = 5000$ ($k = 0.5$). Panels (D) presents estimated odds ratios from the LMM on the true simulated odds ratios for the simulation scenario with $K = 0.02$, $h^2 = 0.5$, $n_{controls} = 8000$, and $n_{cases} = 2000$ ($k = 0.2$). Panel (E) portrays estimated odds ratios from the LMM on the true simulated odds ratios for the simulation scenario with $K = 0.01$, $h^2 = 0.8$, $n_{controls} = 9000$, and $n_{cases} = 1000$ ($k = 0.1$). Panel (F) portrays estimated odds ratios from the LMM on the true simulated odds ratios for the rare variant simulation scenario with $K = 0.01$, $h^2 = 0.05$, $n_{controls} = 8600$, and $n_{cases} = 1400$ ($k = 0.14$). All odds ratios have been reported for the allele that increases the odds of having the disease such that each point is greater than (1, 1). Panels display comparisons from 5,000 simulated true effects generated from the 50 replicates. All panels include the fitted regression line (blue) and $y = x$ line (black) for reference with the key statistics of this regression displayed at the top of each panel.

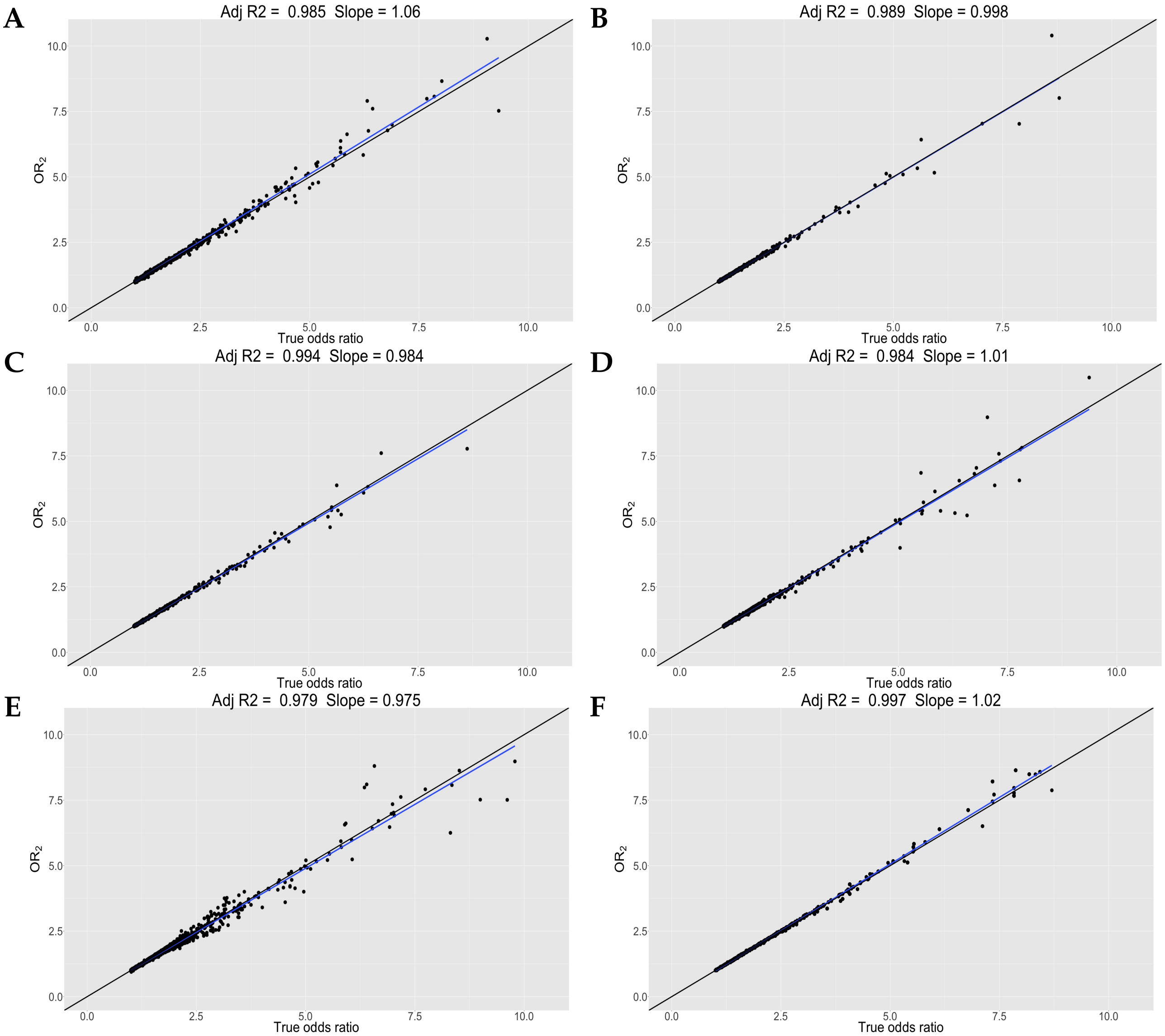

Figure S5 **Performance of odds ratio transformation when using standard error version of approximate odds ratio transformation** OR_2**.** Comparison of estimated odds ratios using OR_2 and reference allele frequencies from the 1000 Genomes Phase 1 Version 3 European sample with true simulated odds ratios from the logistic model simulation and liability threshold model simulation. Panel (A) depicts estimated odds ratios from the LMM on the true simulated odds ratios for the logistic model simulation. Panel (B) depicts estimated odds ratios from the LMM on the true simulated odds ratios for the simulation scenario with $K = 0.1$, $h^2 = 0.5$, $n_{controls} = 5000$, and $n_{cases} = 5000$ ($k = 0.5$). Panel (C) shows estimated odds ratios from the LMM on the true simulated odds ratios for the simulation scenario with $K = 0.05$, $h^2 = 0.5$, $n_{controls} = 5000$, and $n_{cases} = 5000$ ($k = 0.5$). Panel (D) presents estimated odds ratios from the LMM on the true simulated odds ratios for the simulation scenario with $K = 0.02$, $h^2 = 0.5$, $n_{controls} = 8000$, and $n_{cases} = 2000$ ($k = 0.2$). Panel (E) portrays estimated odds ratios from the LMM on the true simulated odds ratios for the simulation scenario with $K = 0.01$, $h^2 = 0.8$, $n_{controls} = 9000$, and $n_{cases} = 1000$ ($k = 0.1$). Panel (F) portrays estimated odds ratios from the LMM on the true simulated odds ratios for the rare variant simulation scenario with $K = 0.01$, $h^2 = 0.05$, $n_{controls} = 8600$, and $n_{cases} = 1400$ ($k = 0.14$). All odds ratios have been reported for the allele that increases the odds of having the disease such that each point is greater than (1, 1). Panels display comparisons from 5,000 simulated true effects generated from the 50 replicates. All panels include the fitted regression line (blue) and $y = x$ line (black) for reference with the key statistics of this regression displayed at the top of each panel.

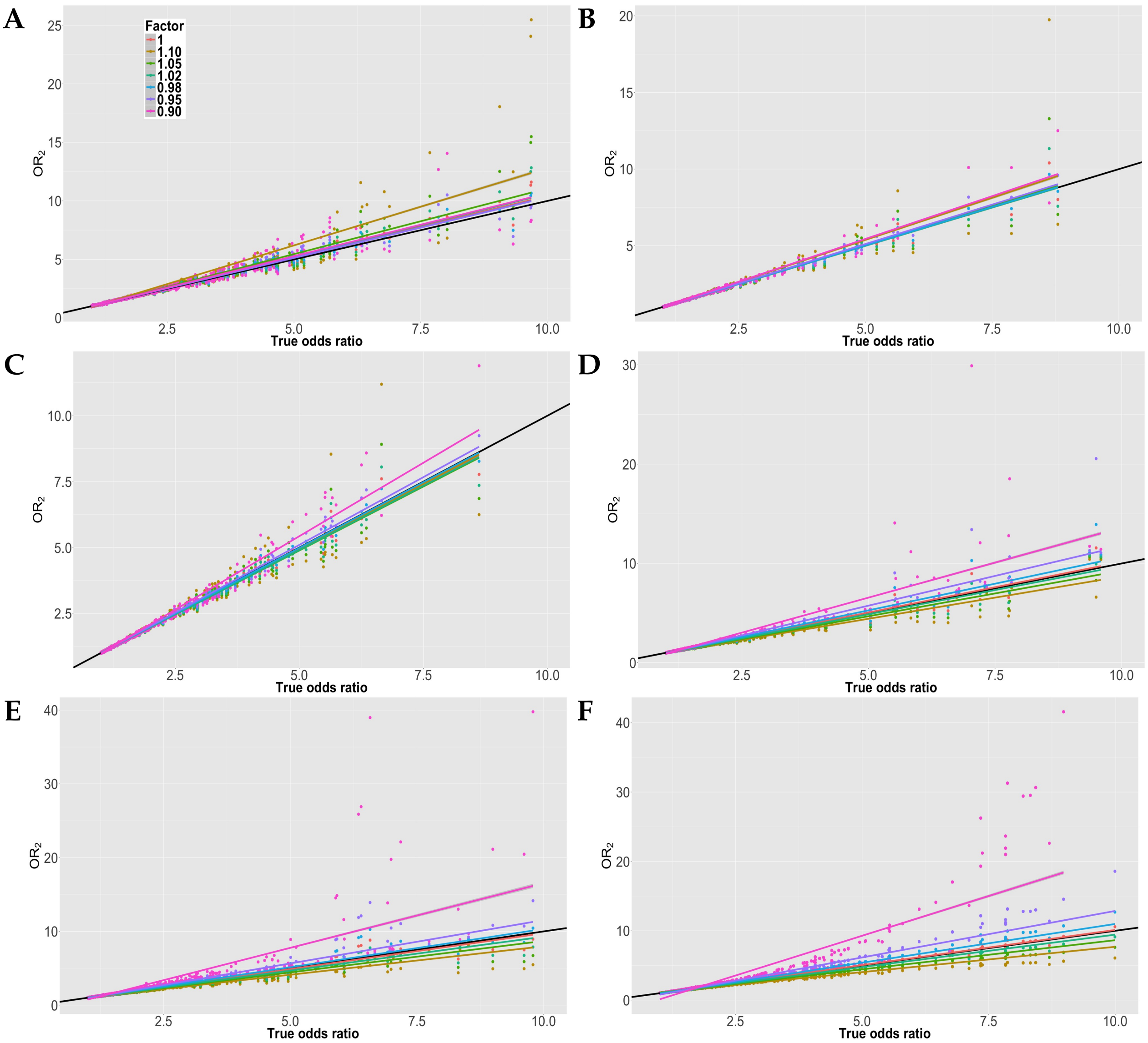

Figure S6 Summary of difference between true odds ratio and estimates from the transformed linear model when k deviates from the true value for simulation scenarios shown in Figure 1. Panels depict the linear regression fit to transformed odds ratio on the true odds ratios for when k was multiplied by the factor in the legend of panel (A), which applies to all panels. Panel (A) depicts estimated odds ratios from the LMM on the true simulated odds ratios for the logistic model simulation. Panel (B) depicts estimated odds ratios from the LMM on the true simulated odds ratios for the simulation scenario with $K = 0.1$, $h^2 = 0.5$, $n_{controls} = 5000$, and $n_{cases} = 5000$ ($k = 0.5$). Panel (C) shows estimated odds ratios from the LMM on the true simulated odds ratios for the simulation scenario with $K = 0.05$, $h^2 = 0.5$, $n_{controls} = 5000$, and $n_{cases} = 5000$ ($k = 0.5$). Panel (D) presents estimated odds ratios from the LMM on the true simulated odds ratios for the simulation scenario with $K = 0.02$, $h^2 = 0.5$, $n_{controls} = 8000$, and $n_{cases} = 2000$ ($k = 0.2$). Panel (E) portrays estimated odds ratios from the LMM on the true simulated odds ratios for the simulation scenario with $K = 0.01$, $h^2 = 0.8$, $n_{controls} = 9000$, and $n_{cases} = 1000$ ($k = 0.1$). Panel (F) portrays estimated odds ratios from the LMM on the true simulated odds ratios for the rare variant simulation scenario with $K = 0.01$, $h^2 = 0.05$, $n_{controls} = 8600$, and $n_{cases} = 1400$ ($k = 0.14$). All odds ratios have been reported for the allele that increases the odds of having the disease such that each point is greater than (1, 1).

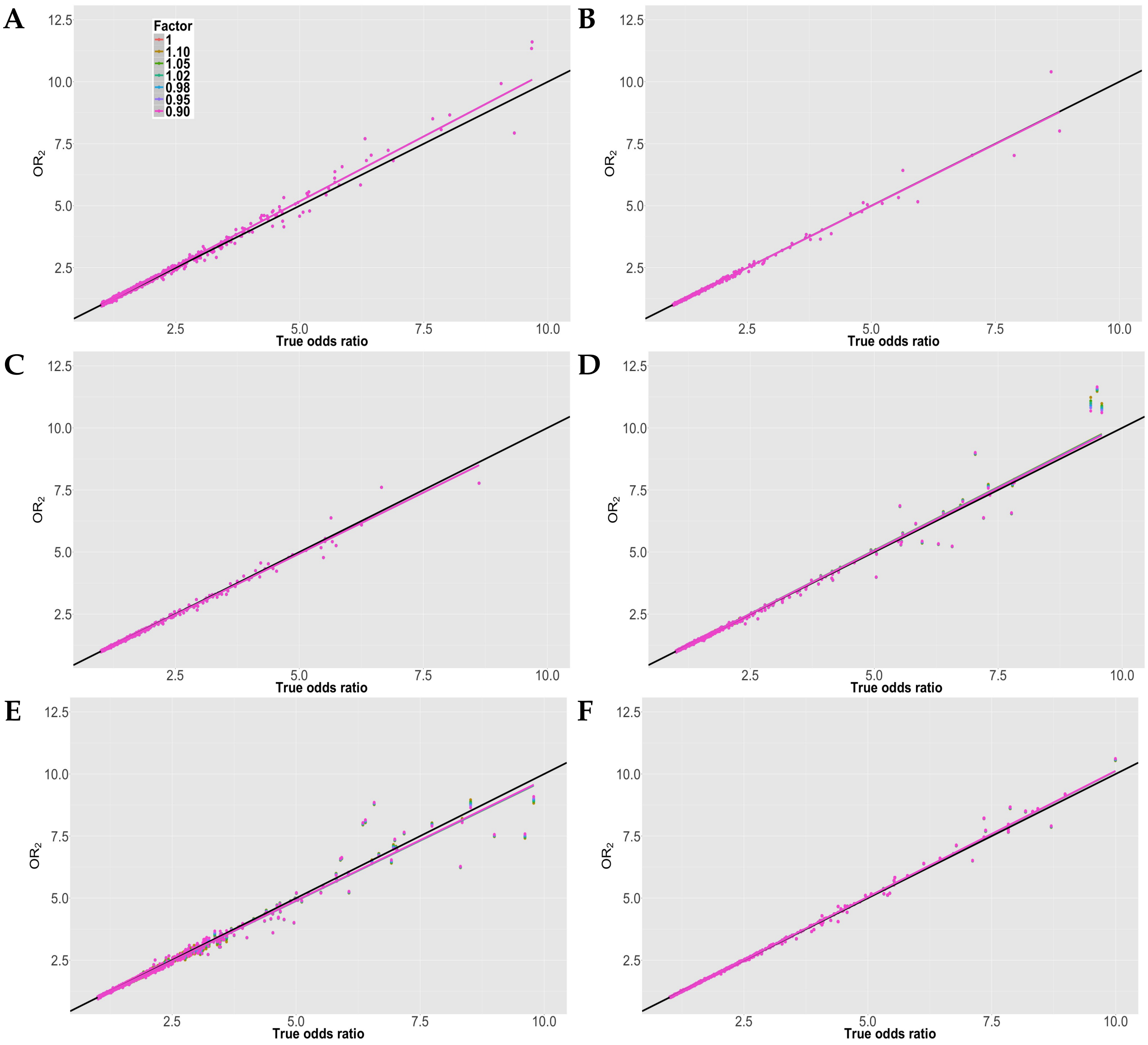

Figure S7 Summary of difference between true odds ratio and estimates from the transformed linear model when p deviates from the true value for simulation scenarios shown in Figure 1. Panels depict the linear regression fit to transformed odds ratio on the true odds ratios for when k was multiplied by the factor in the legend of panel (A), which applies to all panels. Panel (A) depicts estimated odds ratios from the LMM on the true simulated odds ratios for the logistic model simulation. Panel (B) depicts estimated odds ratios from the LMM on the true simulated odds ratios for the simulation scenario with $K = 0.1$, $h^2 = 0.5$, $n_{controls} = 5000$, and $n_{cases} = 5000$ ($k = 0.5$). Panel (C) shows estimated odds ratios from the LMM on the true simulated odds ratios for the simulation scenario with $K = 0.05$, $h^2 = 0.5$, $n_{controls} = 5000$, and $n_{cases} = 5000$ ($k = 0.5$). Panels (D) presents estimated odds ratios from the LMM on the true simulated odds ratios for the simulation scenario with $K = 0.02$, $h^2 = 0.5$, $n_{controls} = 8000$, and $n_{cases} = 2000$ ($k = 0.2$). Panel (E) portrays estimated odds ratios from the LMM on the true simulated odds ratios for the simulation scenario with $K = 0.01$, $h^2 = 0.8$, $n_{controls} = 9000$, and $n_{cases} = 1000$ ($k = 0.1$). Panel (F) portrays estimated odds ratios from the LMM on the true simulated odds ratios for the rare variant simulation scenario with $K = 0.01$, $h^2 = 0.05$, $n_{controls} = 8600$, and $n_{cases} = 1400$ ($k = 0.14$). All odds ratios have been reported for the allele that increases the odds of having the disease such that each point is greater than (1, 1).

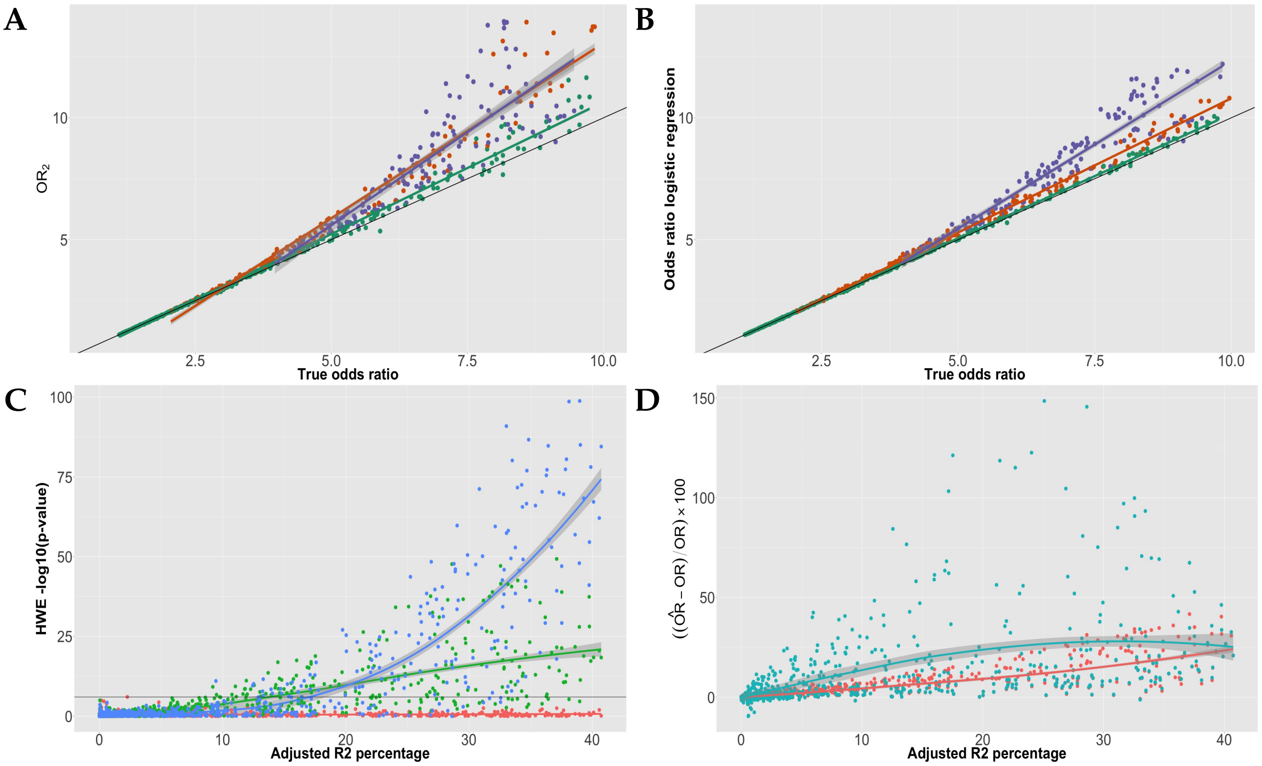

Figure S8 **Single large effect variant simulation without a large covariate effect.** Panels depict results from simulation replicates for a single variant of large effect and a polygenic background such that $h^2 = 0.5$ on the liability scale, a population prevalence $K = 0.01$, and $k = 0.5$. For each scenario, a grid of effects ranging from 0.1 to 1.5 increasing in 0.1 increments was generated with 50 phenotypes generated per effect size in each simulation (750 total). The points in panel (A) show the true odds ratio versus the odds ratio estimated from the transformation (using OR_2) of the estimated effect from a linear model and panel (B) the true odds ratio versus the odds ratio estimated from logistic regression. The colours in panels (A) and (B) represent the variants that have an adjusted R^2 (expressed as a percentage) of (0, 5] (green), (5, 20] (orange) and >20 (purple) (for reference with panel (C)) and the coloured lines the linear model fit to each class of points. The adjusted R^2 was calculated from the regression of the simulated phenotype on the variant of large effect. The black line represents the $y = x$ line. Panel (C) depicts results from the one degree of freedom chi-squared test for Hardy-Weinberg genotype disequilibrium (black line is the 1×10^{-6} value) for each simulated variant in panel (A) for the whole SNP (blue), just cases (green) and just controls (red). The trend lines in panel (C) were fitted using the `loess` method in R. Panel (D) depicts the deviations from the true odds ratio expressed as a proportion (for the points in (A) and (B)) with negative values implying that the odds ratio was underestimated relative to the true value. The colours in panel (D) represent the deviations for the transformed (OR_2) linear regression estimates (aqua) and the logistic regression estimates (red).

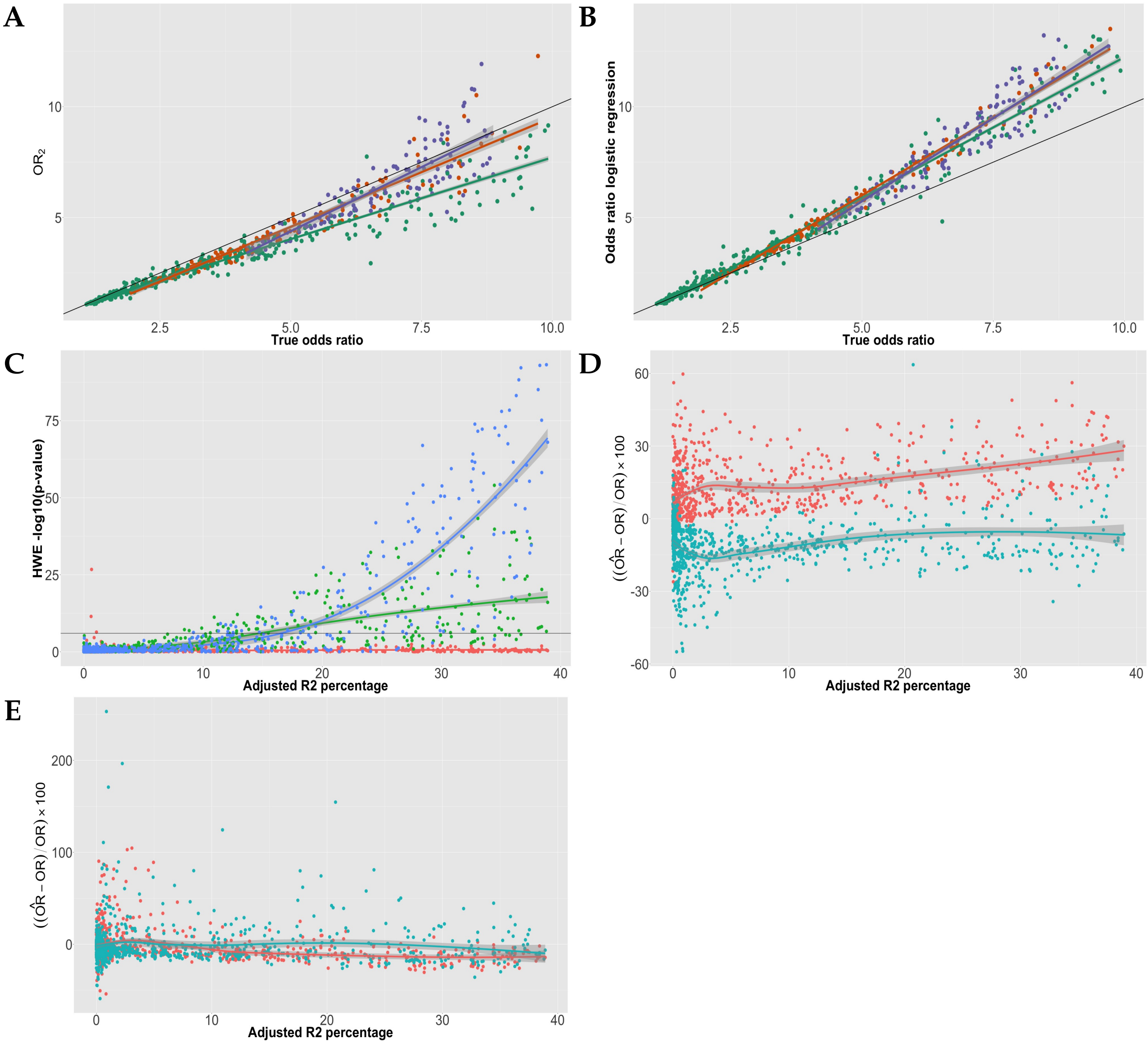

Figure S9 **Single large effect variant simulation with a large covariate effect.** Panels depict results from simulation replicates for a single variant of large effect, a large covariate effect and a polygenic background such that $h^2 = 0.5$ on the liability scale, a population prevalence $K = 0.01$, and $k = 0.5$. For each scenario, a grid of effects ranging from 0.1 to 1.5 increasing in 0.1 increments was generated with 50 phenotypes generated per effect size in each simulation (750 total). The points in panel (A) show the true odds ratio versus the odds ratio estimated the transformation (using OR_2) of the estimated effect from a linear model) and panel (B) the true odds ratio versus the odds ratio estimated from logistic regression. The colours in panels (A) and (B) represent the variants that have an adjusted R^2 (expressed as a percentage) of (0, 5] (green), (5, 20] (orange) and >20 (purple) (for reference with panel (C)) and the coloured lines the linear model fit to each class of points. The adjusted R^2 was calculated from the regression of the simulated phenotype on the variant of large effect. The black line represents the $y = x$ line. Panel (C) depicts results from the one degree of freedom chi-squared test for Hardy-Weinberg genotype disequilibrium (black line is the 1×10^{-6} value) for each simulated variant in panel (A) for the whole SNP (blue), just cases (green) and just controls (red). The trend lines in panel (C) were fitted using the `loess` method in R. Panel (D) depicts the deviations from the true odds ratio expressed as a proportion (for the points in (A) and (B)) with negative values implying that the odds ratio was underestimated relative to the true value. The colours in panel (D) represent the deviations for the transformed (OR_2) linear regression estimates (aqua) and the logistic regression estimates (red). Panel (E) summarises the deviations from the true odds ratio when a meta analysis from the within covariate group estimates was performed using the transformed odds ratio from linear regression and logistic regression.

	[1,2)	[2,3)	[3,4)	[4,5)	[5,6)	[6,7)	[7,8)	[8,9)	[9,10)
No. of variants									
Logistic	4778	128	42	23	12	6	2	1	4
k=0.1	4925	50	9	6	4	0	2	2	0
k=0.05	4895	54	20	11	8	3	0	1	0
k=0.02	4878	67	15	10	10	5	7	0	3
k=0.01	4636	233	59	19	10	10	4	4	2
k=0.01 rare	3126	889	328	271	79	33	50	19	1
Avg. allele frequency									
Logistic	0.417	0.071	0.022	0.018	0.018	0.018	0.011	0.017	0.013
K=0.1	0.439	0.029	0.021	0.024	0.018	0.000	0.010	0.019	0.000
K=0.05	0.441	0.043	0.022	0.020	0.013	0.015	0.000	0.006	0.000
K=0.02	0.442	0.075	0.026	0.022	0.017	0.017	0.018	0.000	0.030
K=0.01	0.448	0.313	0.151	0.018	0.017	0.021	0.014	0.019	0.026
K=0.01 rare	0.004	0.004	0.003	0.004	0.005	0.003	0.004	0.005	0.006

Table S1 Summary the number of causal variants and their average allele frequency for odds ratio bins generated from the results of the simulations scenarios presented in Figures 1 and S3. Odds ratio bins were generated from the results of the logistic model simulation (Logistic), simulation scenario with $K = 0.1$, $h^2 = 0.5$, $n_{controls} = 5000$, and $n_{cases} = 5000$ ($k = 0.5$), simulation scenario with $K = 0.05$, $h^2 = 0.5$, $n_{controls} = 5000$, and $n_{cases} = 5000$ ($k = 0.5$), simulation scenario with $K = 0.02$, $h^2 = 0.5$, $n_{controls} = 8000$, and $n_{cases} = 2000$ ($k = 0.2$), simulation scenario with $K = 0.01$, $h^2 = 0.8$, $n_{controls} = 9000$, and $n_{cases} = 1000$ ($k = 0.1$), and the rare variant simulation scenario with $K = 0.01$, $h^2 = 0.05$, $n_{controls} = 8600$, and $n_{cases} = 1400$ ($k = 0.14$).