Web-based Supplementary Materials for "Using Ad Hoc Methods to Analyze Secondary Traits in Case-Control Association Studies" by Godwin Yung and Xihong Lin

Web Appendix A

Theoretical derivations

Common disease, binary secondary trait

Common disease, continuous secondary trait

Rare disease

Common disease, binary secondary trait

Here, we determine the conditions under which r and r_d are approximately linear functions of $\mathbf{Z}$. First, note that r and r_d are functions of $\mathbf{Z}$ and $\mathbf{G}$ through the conditional means $\mu_D(1)$ and $\mu_D(0)$, which are themselves functions of $\eta = \Phi^{-1}(\mu_D(0)) = \beta_0 + \mathbf{Z}'\boldsymbol{\beta}_Z + \mathbf{G}'\boldsymbol{\beta}_G$. It follows that r and r_d are functions of η and we can write $r(\mathbf{Z},\mathbf{G}) = r(\eta)$ and $r_d(\mathbf{Z},\mathbf{G}) = r_d(\eta)$. We will use the different forms interchangeably. Now consider the second order Taylor series expansions of $r(\eta)$ and $r_d(\eta)$ centered at $\eta_0 = g_D(\kappa)$:

$$\begin{aligned} r(\eta) &= r(\eta_0) + r'(\eta_0)(\eta-\eta_0) + \frac{r''(\eta^*)}{2}(\eta-\eta_0)^2 \\ r_d(\eta) &= r_d(\eta_0) + r_d'(\eta_0)(\eta-\eta_0) + \frac{r_d''(\eta_d^*)}{2}(\eta-\eta_0)^2 \end{aligned}$$

where η^* and η_d^* are some real numbers between η and η_0. One can show that for $g_D(\cdot) =$ logit and $\kappa \in (0.1, 0.5)$,

$$\max_{\eta}\left|\frac{r''(\eta)}{2}\right| \approx \begin{cases} \frac{1}{4}\left(1 - \frac{\kappa}{\widetilde{P}(D=1)}\right)|\beta_Y| & \text{if } \kappa \le \widetilde{P}(D=1) \\ \frac{1}{2}(\kappa - \widetilde{P}(D=1))|\beta_Y| & \text{if } \kappa > \widetilde{P}(D=1) \end{cases}$$

and

$$\left|\frac{r_d''(\eta)}{2}\right| < \frac{1}{20}|\beta_Y| \quad \forall \eta.$$

Similar bounds can be found for $g_D(\cdot) = \Phi^{-1}$ and, in general, any smooth $g_D(\cdot)$. These bounds suggest that if $\boldsymbol{\beta}_G = \mathbf{0}$ and $|\beta_Y|$ and $|\eta - \eta_0|$ are not exceedingly large (i.e., Y and $\mathbf{Z}$ are not strongly associated with D), then the quadratic terms in the Taylor expansions will be small, and the remainders $r(\eta)$ and $r_d(\eta)$ will be approximately linear in $\eta = \beta_0 + \mathbf{Z}'\beta_Z$.

An interesting aside: $r(\eta)$ becomes increasingly linear in η as κ tends to $\widetilde{P}(D = 1)$. In fact, if $\kappa = \widetilde{P}(D = 1)$, then $\pi(0) = \pi(1)$ and it is easy to show from Equation (3) in the article that $r(\cdot)$ is exactly equal to 0. This result reflects the notion that a naïve analysis is valid when the study population is a random sample of the general population. Of course, this condition is not true in the setting of case-control studies.

When $r_0(\cdot)$ and $r_1(\cdot)$ are linear functions of $\mathbf{Z}$, the control-only and case-only analyses can be applied to estimate and make inference on $\boldsymbol{\alpha}_G$. An adjusted analysis is also valid if, in addition, $r_1(\cdot) - r_0(\cdot)$ is a constant. It is easy to show that this required condition is true for $g_D(\cdot) =$ logit and approximately true for $g_D(\cdot) = \Phi^{-1}$ when the disease is common. Specifically, if $g_D(\cdot) =$ logit, then

$$r_1(\mathbf{Z},\mathbf{G}) - r_0(\mathbf{Z},\mathbf{G}) = \beta_Y.$$

Meanwhile, the probit and logit link functions are very close in the mid-range. For η such that $\Phi(\eta) \in (0.2, 0.8)$, the standard normal cumulative distribution can be approximated accurately by a transformed logistic distribution $\Phi(\eta) \approx \text{expit}(\eta/\lambda)$. Popular choices for λ include $\sqrt{3}/\pi$ and 5/8 [Amemiya, 1981]. This approximation implies that for common disease and $g_D(\cdot) = \Phi^{-1}$,

$$r_1(\mathbf{Z}, \mathbf{G}) - r_0(\mathbf{Z}, \mathbf{G}) \approx \beta_Y/\lambda.$$

For $g_D(p) = \log(-\log(1-p))$, it can be shown, by taking a second order Taylor series expansion of

$$T(\eta + \beta_Y) = \log\left\{\frac{g_D^{-1}(\eta + \beta_Y)}{1 - g_D^{-1}(\eta + \beta_Y)}\right\}$$

centered at η, that

$$r_1(\mathbf{Z}, \mathbf{G}) - r_0(\mathbf{Z}, \mathbf{G}) = T(\eta + \beta_Y) - T(\eta) \approx T'(\eta^*)\beta_Y + 0.5T''(\eta^*)\beta_Y^2 \approx 1.3\beta_Y + 0.22\beta_Y^2.$$

Common disease, continuous secondary trait

Here, we derive Equations (7) and (8) from the article and provide the closed form expressions for $\widetilde{\mu}_Y$ and $\widetilde{\sigma}^2$. We then determine the conditions under which $r(\cdot)$ and $r_d(\cdot)$ are approximately linear in $\mathbf{Z}$ and $\mathbf{X}$. First, suppose that θ is a parameter in $\mathbb{R}$ and $Y \sim N(\mu_Y + \theta\sigma^2, \sigma^2)$. Our interest is in calculating $E(Y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, D, S = 1, \theta = 0)$ and $Var(Y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, D, S = 1, \theta = 0)$, but since $\widetilde{P}(Y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, D) = P(Y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, D)$, it suffices to calculate the mean and variance of $Y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, D, \theta = 0$. With that in mind, assume $g_D = \Phi^{-1}$ and define D^* to be the random variable such that $D^* = g_D(\mu_D(Y)) + \varepsilon$, $\varepsilon \sim N(0, 1)$. Then $P(D^* > 0|\mathbf{Z}, \mathbf{G}, Y) = P(D = 1|\mathbf{Z}, \mathbf{G}, Y)$. Furthermore,

$$\begin{aligned}
P(D = 1|\mathbf{Z}, \mathbf{G}, \mathbf{X}, \theta) &= \int P(D = 1|\mathbf{Z}, \mathbf{G}, \mathbf{X}, \theta, y)P(y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, \theta)dy \\
&= \int P(D^* > 0|\mathbf{Z}, \mathbf{G}, \mathbf{X}, \theta, y)P(y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, \theta)dy && D|\mathbf{Z}, \mathbf{G}, y \perp\!\!\!\perp \mathbf{X}, \theta \\
&= P(D^* > 0|\mathbf{Z}, \mathbf{G}, \mathbf{X}, \theta) \\
&= P(\varepsilon^* > -g_D(\mu_D(0))) && \varepsilon^* \sim N\left((\mu_Y + \theta\sigma^2)\beta_Y, \sigma^2\beta_Y^2 + 1\right) \\
&= \Phi(f(\theta))
\end{aligned}$$

where

$$f(\theta) = \frac{g_D(\mu_D(\mu_Y + \theta\sigma^2))}{\sqrt{\sigma^2\beta_Y^2 + 1}}.$$

We can now calculate the first two moments of $Y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, D, \theta = 0$ via its moment generating function:

$$\begin{aligned}
E(e^{tY}|\mathbf{X}, \mathbf{G}, \mathbf{Z}, D = d, \theta = 0) &= \exp\left(t\mu_Y + \frac{t^2\sigma^2}{2}\right)\left\{\frac{\Phi(f(t))}{\Phi(f(0))}\right\}^d \left\{\frac{1-\Phi(f(t))}{1-\Phi(f(0))}\right\}^{1-d} \\
E(Y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, D = d, \theta = 0) &= \mu_Y + (-1)^{1-d} \cdot c \cdot \frac{\phi(f(0))}{\{\Phi(f(0))\}^d\{1-\Phi(f(0))\}^{1-d}} \\
E(Y^2|\mathbf{X}, \mathbf{G}, \mathbf{Z}, D = d, \theta = 0) &= \sigma^2 + (-1)^{1-d} \cdot c^2 \cdot \frac{\phi'(f(0))}{\{\Phi(f(0))\}^d\{1-\Phi(f(0))\}^{1-d}} \\
&\quad + \mu_Y^2 + 2\mu_Y(-1)^{1-d} \cdot c \cdot \frac{\phi(f(0))}{\{\Phi(f(0))\}^d\{1-\Phi(f(0))\}^{1-d}}
\end{aligned}$$

The variance follows immediately:

$$Var(Y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, D = d, \theta = 0) = \sigma^2 + c^2\left(\frac{(-1)^{1-d} \cdot \phi'(f(0))}{\{\Phi(f(0))\}^d\{1-\Phi(f(0))\}^{1-d}} - \frac{\phi(f(0))^2}{\{\Phi(f(0))\}^{2d}\{1-\Phi(f(0))\}^{2(1-d)}}\right)$$

Letting η denote $f(0)$ gives us Equations (7) and (8).

Next, to calculate $\widetilde{\mu}_Y$ and $\widetilde{\sigma}^2$, note that

$$\widetilde{P}(D|\mathbf{Z}, \mathbf{G}, \mathbf{X}, \theta) = \frac{P(S = 1|D) \cdot P(D|\mathbf{Z}, \mathbf{G}, \mathbf{X}, \theta)}{\sum_{d=0}^1 P(S = 1|D = d) \cdot P(D = d|\mathbf{Z}, \mathbf{G}, \mathbf{X}, \theta)}.$$

Therefore,

$$\begin{aligned}
E(Y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, S = 1, \theta = 0) &= \frac{\sum_{d=0}^1 E(Y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, D = d, \theta = 0) \cdot P(S = 1|D = d) \cdot P(D = d|\mathbf{Z}, \mathbf{G}, \mathbf{X}, \theta = 0)}{\sum_{d=0}^1 P(S = 1|D = d) \cdot P(D = d|\mathbf{Z}, \mathbf{G}, \mathbf{X}, \theta = 0)} \\
&= \mu_Y + c \cdot \phi(\eta) \cdot g(\mathbf{Z}, \mathbf{G}, \mathbf{X}) \\
&= \mu_Y + r(\mathbf{Z}, \mathbf{G}, \mathbf{X}) \\
Var(Y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, S = 1, \theta = 0) &= E(Var(Y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, D, \theta = 0)) + Var(E(Y|\mathbf{X}, \mathbf{G}, \mathbf{Z}, D, \theta = 0)) \\
&= \sigma^2 + c^2 \cdot \left\{\phi'(\eta) \cdot g(\mathbf{Z}, \mathbf{G}, \mathbf{X}) - \phi(\eta)^2 \cdot g(\mathbf{Z}, \mathbf{G}, \mathbf{X})^2\right\} \\
&= \sigma^2 + s(\mathbf{Z}, \mathbf{G}, \mathbf{X})
\end{aligned}$$

where

$$g(\mathbf{Z},\mathbf{G},\mathbf{X}) = \frac{P(S=1|D=1) - P(S=1|D=0)}{\sum_{d=0}^{1} P(S=1|D=d)\cdot P(D=d|\mathbf{Z},\mathbf{G},\mathbf{X},\theta=0)}.$$

Finally, we determine the conditions under which $r(\cdot)$ and $r_d(\cdot)$ are approximately linear in $\mathbf{Z}$ and $\mathbf{X}$, and s and s_d are approximately constants. We begin by again noting that the all remainders are a function of η, which is itself a linear function of $\mathbf{Z}$, $\mathbf{G}$, and $\mathbf{X}$. Thus, we can write $r(\mathbf{Z},\mathbf{G},\mathbf{X}) = r(\eta)$, $r_d(\mathbf{Z},\mathbf{G},\mathbf{X}) = r_d(\eta)$, $s(\mathbf{Z},\mathbf{G},\mathbf{X}) = s(\eta)$, and $s_d(\mathbf{Z},\mathbf{G},\mathbf{X}) = s_d(\eta)$. If $\boldsymbol{\alpha}_G = \boldsymbol{\beta}_G = \mathbf{0}$, then the remainders are functions of $\mathbf{Z}$ and $\mathbf{X}$ alone. Meanwhile, consider the second and first order order Taylor series expansions of $r_d(\eta)$ and $s_d(\eta)$ centered at $\eta_0 = g_D(\kappa)/\sqrt{\sigma^2\beta_Y^2+1}$. One can show that in these expansions the quadratic and linear coefficients are bounded:

$$\left|\frac{r_d''(\eta)}{2}\right| < \frac{3}{20}|c|$$

and

$$|s_d'(\eta)| < \frac{3}{10}c^2$$

for all η and $d = 0,1$. Similar bounds can be derived for $r(\eta)$ and $s(\eta)$. Therefore, if $\boldsymbol{\alpha}_G = \boldsymbol{\beta}_G = \mathbf{0}$ and $|\beta_Y|$ and $|\eta - \eta_0|$ are not exceedingly large (i.e., Y and $\mathbf{Z}$ are not strongly associated with D and $\mathbf{X}$ is not strongly associated with Y), then the quadratic and linear terms in the Taylor expansion of $r(\eta)$, $r_d(\eta)$, $s(\eta)$, and $s_d(\eta)$ will be small, $r(\eta)$ and $r_d(\eta)$ will be approximately linear in η—hence in $\mathbf{X}$ and $\mathbf{Z}$—and $s(\eta)$ and $s_d(\eta)$ we be approximately constant.

An adjusted analysis is unbiased if, in addition, $\boldsymbol{\alpha}_{Z0}^{**} = \boldsymbol{\alpha}_{Z1}^{**}$ and $\boldsymbol{\alpha}_{X0}^{**} = \boldsymbol{\alpha}_{X1}^{**}$ or, equivalently, $r_1(\cdot) - r_0(\cdot)$ is a constant. It is easy to show that this required condition is approximately true for common disease by using the logit approximation for the probit:

$$r_1(\mathbf{X},\mathbf{G},\mathbf{Z}) - r_0(\mathbf{X},\mathbf{G},\mathbf{Z}) \approx c/\lambda$$

While $s_0(\eta)$ is generally not equal to $s_1(\eta)$, in our simulations, the difference between the sample variance of the case-only and control-only analyses with pooled covariates seemed to be small enough for inference to be approximately correct.

Rare disease

Here, we derive the theoretical bias for the case-only analysis with pooled covariates when the disease is rare and $g_D = \Phi^{-1}$. If Y is binary, then

$$\begin{aligned}
\lim_{\eta\to-\infty} r_1'(\eta) &= \lim_{\eta\to-\infty} \frac{\phi(\eta+\beta_Y)}{\Phi(\eta+\beta_Y)} - \frac{\phi(\eta)}{\Phi(\eta)} \\
&= \lim_{\eta\to-\infty} \frac{\phi(\eta)}{\Phi(\eta)} + \frac{\phi'(\eta^*)\Phi(\eta^*) - \phi(\eta^*)^2}{\Phi(\eta^*)^2}\beta_Y - \frac{\phi(\eta)}{\Phi(\eta)} && \eta^* \text{ between } \eta \text{ and } \eta+\beta_Y \\
&= \beta_Y \cdot \lim_{\eta\to-\infty} \frac{\phi'(\eta)\Phi(\eta) - \phi(\eta)^2}{\Phi(\eta)^2} \\
&= \beta_Y \cdot \lim_{\eta\to-\infty} \frac{\eta\phi(\eta) + (\eta^2-1)\Phi(\eta)}{2\Phi(\eta)} && \text{L'Hopital's rule} \\
&= \beta_Y \cdot \lim_{\eta\to-\infty} \frac{\eta\Phi(\eta)}{\phi(\eta)} && \text{L'Hopital's rule} \\
&= -\beta_Y
\end{aligned}$$

and

$$\begin{aligned}
\lim_{\eta\to-\infty} r_1''(\eta) &= \lim_{\eta\to-\infty} \frac{d}{d(\eta+\beta_Y)}\frac{\phi(\eta+\beta_Y)}{\Phi(\eta+\beta_Y)} - \frac{d}{d(\eta)}\frac{\phi(\eta)}{\Phi(\eta)} \\
&= (-1) - (-1) \\
&= 0.
\end{aligned}$$

It follows that

$$\lim_{\eta\to-\infty} r_1(\eta) = r_1(\Phi^{-1}(\kappa)) - \beta_Y(\eta - \Phi^{-1}(\kappa)),$$

or equivalently,

$$\lim_{\kappa\to 0} r_1(\mathbf{Z},\mathbf{G}) = \left\{ r_1(\Phi^{-1}(\kappa)) + \beta_Y(\Phi^{-1}(\kappa) - \beta_0) \right\} - \mathbf{Z}'\beta_Y\boldsymbol{\beta}_Z - \mathbf{G}'\beta_Y\boldsymbol{\beta}_G.$$

If instead Y is continuous, then because $\lim_{\eta\to-\infty} \frac{\phi(\eta)}{\eta\Phi(\eta)} = -1$,

$$r_1(\mathbf{Z},\mathbf{G},\mathbf{X}) = c\frac{\phi(\eta)}{\Phi(\eta)} \approx -c\eta.$$

Meanwhile,

$$\lim_{\kappa\to 0} s_1(\mathbf{Z},\mathbf{G},\mathbf{X}) = \lim_{\eta\to-\infty} c^2 \cdot \left\{ \frac{\phi'(\eta)\Phi(\eta) - \phi(\eta)^2}{\Phi(\eta)^2} \right\} = -c^2.$$

Web Appendix B

Simulation study with binary secondary trait Y

Type I error rates

Bias

Power

In the paper, we presented simulation results for continuous Y. Here, we do the same but for binary Y. The simulation setting for binary Y is similar to that for continuous Y, the only difference being instead of sampling Y_i from a normal distribution, Y_i is sampled as a Bernoulli random variable with probability of success $\text{expit}(\alpha_0 + X_{2i}\alpha_{X2} + X_{3i}\alpha_{X3} + G_i\alpha_G)$ with $\alpha_{X2} = \alpha_{X3} = 0.2$, $\alpha_G \in \{0, \log(1.7)/2, \log(1.7)\}$, and α_0 chosen so that the secondary trait Y_i has a prevalence of 0.10 in the population. In order to estimate type I error rate (power) accurately, a total of 10^6 (10^4) replicate data sets were simulated for each scenario.

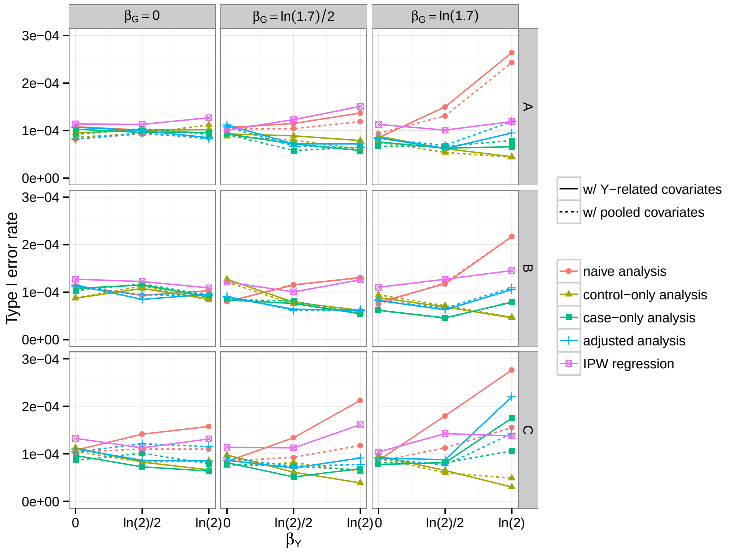

Figure 1: Empirical type I error rates for testing genetic associations with a binary secondary trait, at genome-wide $\alpha = 10^{-4}$ level and across scenarios with different combinations of β_Y, β_G, γ_1 and β_{Z1}. Nine methods are compared here. Each method takes either a naïve, control-only, case-only, adjusted, or IPW approach, and adjusts for covariates related to Y or covariates related to (Y, D). The disease is assumed to be common (10% prevalence) and to follow a logistic model (g_D = logit). In row **A**, covariate Z_1 is assumed to be associated with G but not with D ($\gamma_1 = \ln 1.7$, $\beta_{Z1} = 0$). In row **B**, Z_1 is associated with D but not with G ($\gamma_1 = 0$, $\beta_{Z1} = \ln 1.7$). In row **C**, Z_1 is a confounder of the association between G and D ($\gamma_1 = \beta_{Z1} = \ln 1.7$).

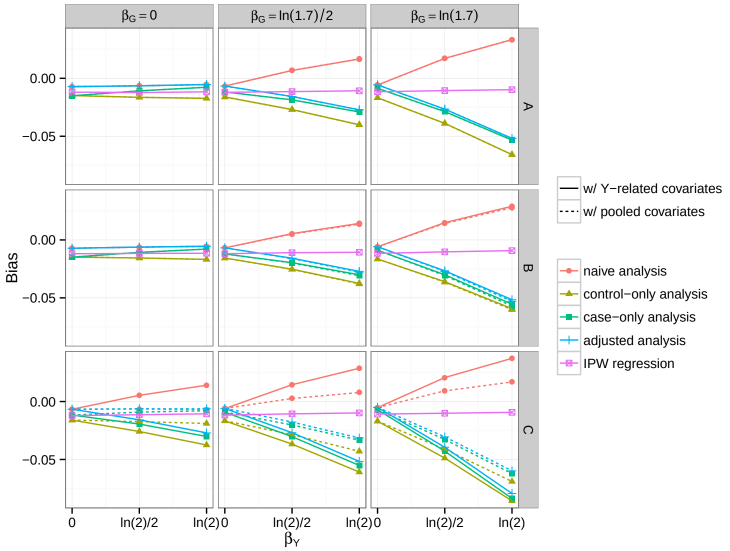

Figure 2: Empirical bias for the estimated genetic effect $\widehat{\alpha}_G$ on a binary secondary trait, across null scenarios ($\alpha_G = 0$) with different combinations of β_Y, β_G, γ_1 and β_{Z1}. Nine methods are compared here. Each method takes either a naïve, control-only, case-only, adjusted, or IPW approach, and adjusts for covariates related to Y or covariates related to (Y, D). The disease is assumed to be common (10% prevalence) and to follow a logistic model (g_D = logit). In row **A**, covariate Z_1 is assumed to be associated with G, but not with D ($\gamma_1 = \ln 1.7$, $\beta_{Z1} = 0$). In row **B**, Z_1 is associated with D, but not with G ($\gamma_1 = 0$, $\beta_{Z1} = \ln 1.7$). In row **C**, Z_1 is a confounder of the association between G and D ($\gamma_1 = \beta_{Z1} = \ln 1.7$).

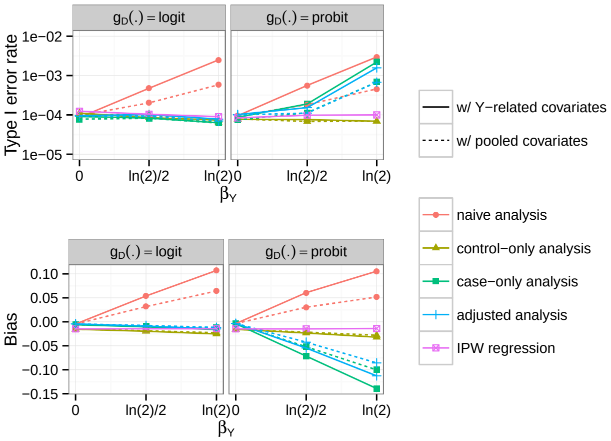

Figure 3: Empirical type I error rates and bias for testing and estimating genetic associations with a binary secondary trait, at genome-wide $\alpha = 10^{-4}$ level and across null scenarios ($\alpha_G = 0$) with different combinations of β_Y and link function g_D for the disease model. Nine methods are compared here. Each method takes either a naïve, control-only, case-only, adjusted, or IPW approach, and adjusts for covariates related to Y or covariates related to (Y, D). The disease is assumed to be rare (1% prevalence) and to follow either a logistic or probit model ($g_D = \text{logit}$ or Φ^{-1}). G is assumed to be associated with D ($\beta_G = \ln 1.7$). Z_1 is assumed to be a confounder of the association between G and D ($\gamma_1 = \beta_{Z1} = \ln 1.7$). The scenarios with a logistic disease model (left column) are the same as the scenarios in the bottom right plots of Figures 1 and 2, except here the disease is rare, not common.

Table 1: Empirical bias for the estimated genetic of $\widehat{\alpha}_G$ based on the case-only analysis with pooled covariates, across null scenarios ($\alpha_G = 0$), and over varying levels of disease prevalence κ. The disease is assumed to follow a probit model. G is assumed to be associated with D ($\beta_G = \ln 1.7$). Z_1 is assumed to be a confounder of the association between G and D ($\gamma_1 = \beta_{Z1} = \ln 1.7$).

	$\beta_Y = 0$	$\beta_Y = \frac{\ln(2)}{2}$	$\beta_Y = \ln(2)$
Average			
$\kappa = 0.10$	-0.011	-0.030	-0.064
$\kappa = 0.01$	-0.005	-0.053	-0.101
$\kappa = 0.001$	-0.003	-0.055	-0.102
Theoretical	0.000	-0.056	-0.112

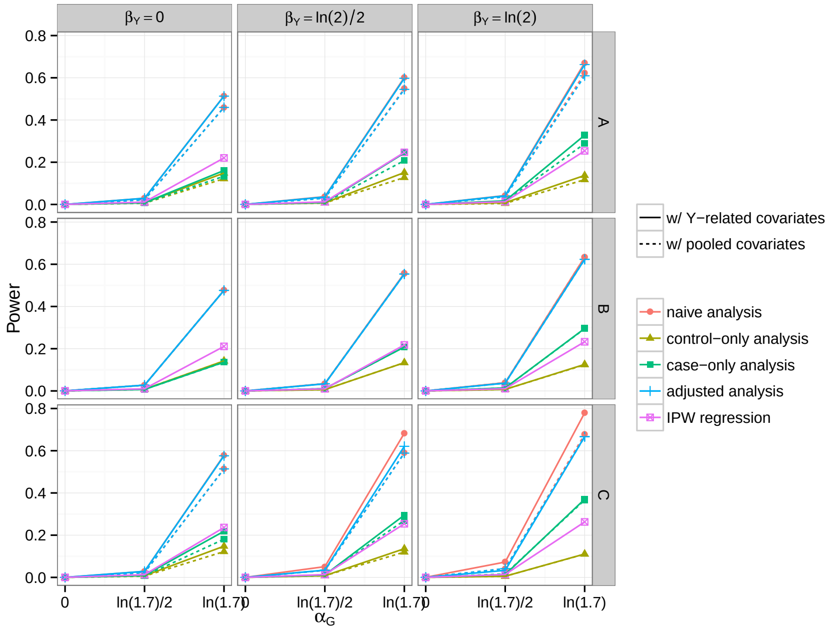

Figure 4: Power for testing genetic associations with a binary secondary trait, at genome-wide $\alpha = 10^{-4}$ level and across scenarios with different combinations of α_G, β_Y, γ_1, and β_{Z1}. Nine methods are compared here. Each method takes either a naïve, control-only, case-only, adjusted, or IPW approach, and adjusts for covariates related to Y or covariates related to (Y, D). The disease is assumed to be common (10% prevalence) and to follow a logistic model ($g_D = \text{logit}$). In row **A**, covariate Z_1 is assumed to be associated with G but not with D ($\gamma_1 = \ln 1.7$, $\beta_{Z1} = 0$). In row **B**, Z_1 is associated with D but not with G ($\gamma_1 = 0$, $\beta_{Z1} = \ln 1.7$). In row **C**, Z_1 is a confounder of the association between G and D ($\gamma_1 = \beta_{Z1} = \ln 1.7$).

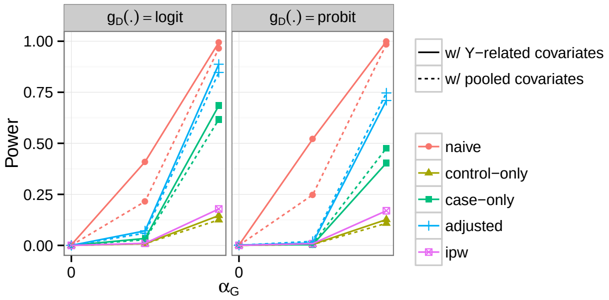

Figure 5: Power for testing genetic associations with a binary secondary trait, at genome-wide $\alpha = 10^{-4}$ level and across scenarios with different combinations of α_G and link function $g_D(\cdot)$ for the disease model. Nine methods are compared here. Each method takes either a naïve, control-only, case-only, adjusted, or IPW approach, and adjusts for covariates related to Y or covariates related to (Y, D). The disease is assumed to be rare (1% prevalence) and to follow either a logistic or probit model ($g_D(\cdot) = \text{logit}$ or Φ^{-1}). G is assumed to be associated with D ($\beta_G = \ln 1.7$). Z_1 is assumed to be a confounder of the association between G and D ($\gamma_1 = \beta_{Z1} = \ln 1.7$).

Web Appendix C

Simulation study with continuous secondary trait Y

Type I error rates

Bias

Power

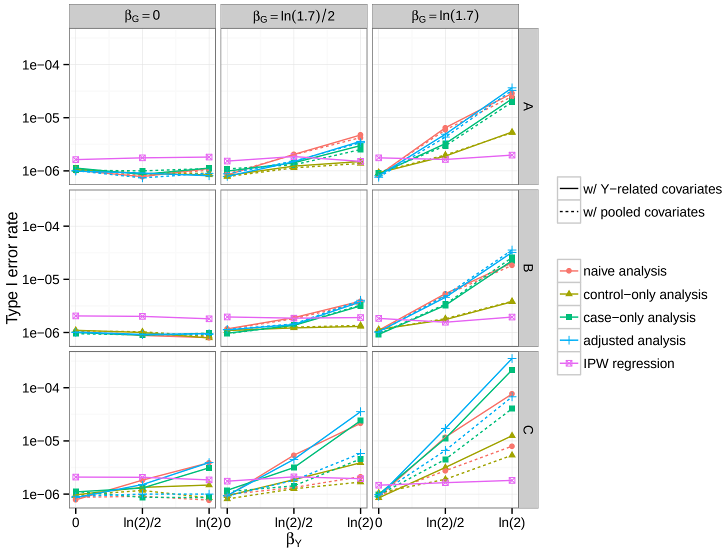

Figure 6: Empirical type I error rates for testing genetic associations with a continuous secondary trait, at genome-wide $\alpha = 10^{-6}$ level and across scenarios with different combinations of β_Y, β_G, γ_1 and β_{Z1}. Nine methods are compared here. Each method takes either a naïve, control-only, case-only, adjusted, or IPW approach, and adjusts for covariates related to Y or covariates related to (Y, D). The disease is assumed to be common (10% prevalence) and to follow a logistic model ($g_D = \text{logit}$). In row **A**, covariate Z_1 is assumed to be associated with G but not with D ($\gamma_1 = \ln 1.7$, $\beta_{Z1} = 0$). In row **B**, Z_1 is associated with D but not with G ($\gamma_1 = 0$, $\beta_{Z1} = \ln 1.7$). In row **C**, Z_1 is a confounder of the association between G and D ($\gamma_1 = \beta_{Z1} = \ln 1.7$).

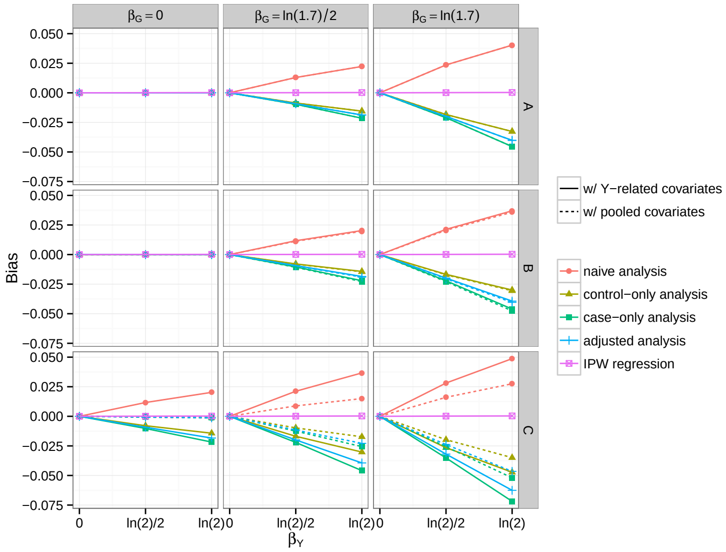

Figure 7: Empirical bias for the estimated genetic effect $\widehat{\alpha}_G$ on a continuous secondary trait, across null scenarios ($\alpha_G = 0$) with different combinations of β_Y, β_G, γ_1 and β_{Z1}. Nine methods are compared here. Each method takes either a naïve, control-only, case-only, adjusted, or IPW approach, and adjusts for covariates related to Y or covariates related to (Y, D). The disease is assumed to be common (10% prevalence) and to follow a logistic model ($g_D = \text{logit}$). In row **A**, covariate Z_1 is assumed to be associated with G, but not with D ($\gamma_1 = \ln 1.7$, $\beta_{Z1} = 0$). In row **B**, Z_1 is associated with D, but not with G ($\gamma_1 = 0$, $\beta_{Z1} = \ln 1.7$). In row **C**, Z_1 is a confounder of the association between G and D ($\gamma_1 = \beta_{Z1} = \ln 1.7$).

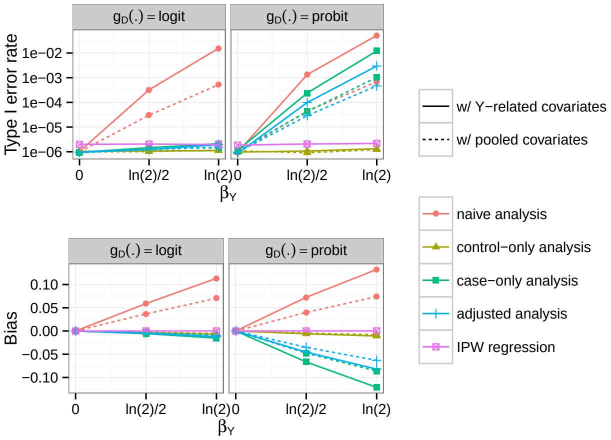

Figure 8: Empirical type I error rates and bias for testing and estimating genetic associations with a continuous secondary trait, at genome-wide $\alpha = 10^{-6}$ level and across null scenarios ($\alpha_G = 0$) with different combinations of β_Y and link function g_D for the disease model. Nine methods are compared here. Each method takes either a naïve, control-only, case-only, adjusted, or IPW approach, and adjusts for covariates related to Y or covariates related to (Y, D). The disease is assumed to be rare (1% prevalence) and to follow either a logistic or probit model ($g_D = \text{logit}$ or Φ^{-1}). G is assumed to be associated with D ($\beta_G = \ln 1.7$). Z_1 is assumed to be a confounder of the association between G and D ($\gamma_1 = \beta_{Z1} = \ln 1.7$). The scenarios with a logistic disease model (left column) are the same as the scenarios in the bottom right plots of Figures **??** and **??**, except here the disease is rare, not common.

Table 2: Empirical bias for the estimated genetic effect $\widehat{\alpha}_G$ based on the case-only analysis with pooled covariates, across null scenarios ($\alpha_G = 0$), and over varying levels of disease prevalence κ. The disease is assumed to follow a probit model. G is assumed to be associated with D ($\beta_G = \ln 1.7$). Z_1 is assumed to be a confounder of the association between G and D ($\gamma_1 = \beta_{Z1} = \ln 1.7$).

	$\beta_Y = 0$	$\beta_Y = \frac{\ln(2)}{2}$	$\beta_Y = \ln(2)$
Empirical			
$\kappa = 0.10$	0.000	-0.026	-0.052
$\kappa = 0.01$	0.000	-0.048	-0.086
Theoretical	0.000	-0.054	-0.098

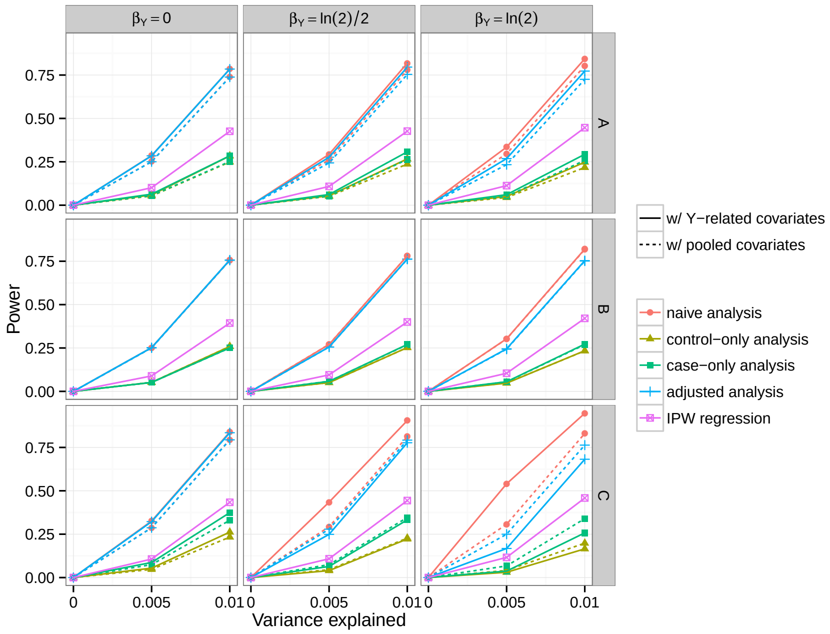

Figure 9: Power for testing genetic associations with a continuous secondary trait, at genome-wide $\alpha = 10^{-4}$ level and across scenarios with different combinations of percent of variance in Y explained by G (r^2_{YG}), β_Y, γ_1, and β_{Z1}. Nine methods are compared here. Each method takes either a naïve, control-only, case-only, adjusted, or IPW approach, and adjusts for covariates related to Y or covariates related to (Y, D). The disease is assumed to be common (10% prevalence) and to follow a logistic model (g_D = logit). In row **A**, covariate Z_1 is assumed to be associated with G but not with D ($\gamma_1 = \ln 1.7$, $\beta_{Z1} = 0$). In row **B**, Z_1 is associated with D but not with G ($\gamma_1 = 0$, $\beta_{Z1} = \ln 1.7$). In row **C**, Z_1 is a confounder of the association between G and D ($\gamma_1 = \beta_{Z1} = \ln 1.7$).

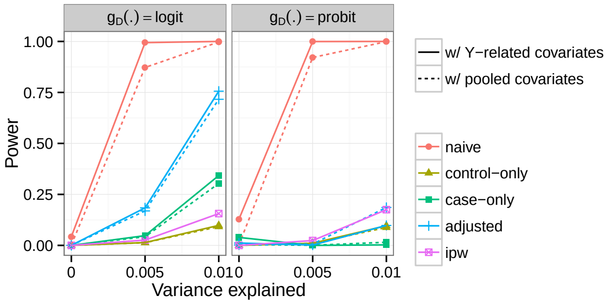

Figure 10: Power for testing genetic associations with a continuous secondary trait, at genome-wide $\alpha = 10^{-4}$ level and across scenarios with different combinations of percent of variance in Y explained by G (r^2_{YG}) and link function $g_D(\cdot)$ for the disease model. Nine methods are compared here. Each method takes either a naïve, control-only, case-only, adjusted, or IPW approach, and adjusts for covariates related to Y or covariates related to (Y, D). The disease is assumed to be rare (1% prevalence) and to follow either a logistic or probit model ($g_D(\cdot)$ = logit or Φ^{-1}). G is assumed to be associated with D ($\beta_G = \ln 1.7$). Z_1 is assumed to be a confounder of the association between G and D ($\gamma_1 = \beta_{Z1} = \ln 1.7$).

Web Appendix D

Analysis of lung cancer data

Calculation of inverse probability weights

Testing for gene-environment interactions (model and results)

Manhattan plots

Top 10 SNPs

Top 10 novel SNPs

Following the article's notation, let D denote the disease status (1=case, 0=control) and S indicate with the values 1 versus 0 whether or not an individual is included in the case-control study. Let E denote ever-smoker status (1=ever-smoker, 0=never-smoker). Then by Bayes' theorem, the inverse-probability-of-sampling weight for case and control ever-smokers is given by

$$P(S=1|D=d,E=1)=\frac{P(D=d|S=1,E=1)P(S=1|E=1)}{P(D=d|E=1)}\propto\frac{P(D=d|S=1)}{P(D=d|E=1)}.$$

We can calculate $P(D=1|E=1)$, the prevalence of lung cancer amongst ever-smokers in Massachusetts, using national and Massachusetts-specific statistics. Specifically, it is estimated that 10-15% of all lung cancers in the U.S. arise in never smokers [Samet et al., 2011]; 0.0745% of the Massachusetts general population have lung cancer [Ter-Minassian et al., 2012]; and 44% of adults in Massachusetts are ever-smokers (Massachusetts Department of Public Health, 2013). It follows by another application of Bayes' rule that

$$\begin{aligned}P(D=1|E=1) &= \frac{P(E=1|D=1)P(D=1)}{P(E=1|D=1)P(D=1)+P(E=1|D=0)P(D=0)}\\ &= \frac{P(E=1|D=1)P(D=1)}{P(E=1|D=1)P(D=1)+P(D=0|E=1)P(E=1)}\\ &= \frac{(0.875)(0.000745)}{(0.875)(0.000745)+P(D=0|E=1)(0.44)}\end{aligned}$$

and $P(D=1|E=1)=0.00148$.

To test for interactions between SNPs and smoking behavior, we fitted for each of the 513,271 SNPs the following model for lung cancer risk

$$\text{logit}(\mu_D(Y))=\beta_0+\mathbf{Z}'\boldsymbol{\beta}_Z+G\beta_G+Y\beta_Y+GY\beta_{GY}.$$

We found that the naïve, case-only, and adjusted estimates for the genetic effect of smoking behavior tended to deviate from the control-only estimate as $\widehat{\beta}_{GY}$ deviated from 0 (Figure 11). Therefore, it would be inappropriate to use ad hoc methods other than the control-only analysis when there is evidence for G-E interaction.

We also found that SNPs identified by the naïve or adjusted analysis tend not to modify the effect of smoking behavior on lung cancer risk. In contrast, many of the SNPs identified by the control-only or IPW analysis had moderate to strong evidence of an interaction with smoking behavior (Tables 3 and 4). This difference between SNPs identified by the naïve or adjusted analysis and SNPs identified by the control-only or IPW analysis is likely due to the methods having more power to detect different sets of SNPs. It explains why we observed in Figure 5 of the article relatively little overlap, and why some previously known genes were only identified by the adjusted analysis (*HSD17B2* and *SLC9A2*) while other known genes were identified by the control-only and IPW analyses but not by the adjusted analysis (*CDH18*).

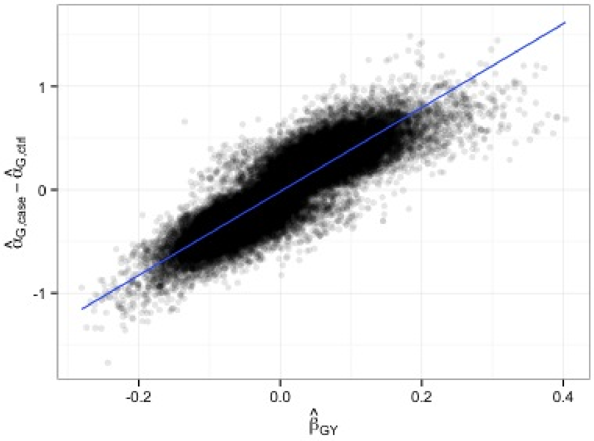

Figure 11: Top 50k SNPs from IPW regression. Observed difference between case-only and control-only estimates has a significant tendency to increase as the difference in effect of smoking behavior on lung cancer risk between adjacent genotypes (2 vs 1, 1 vs 0) increases (slope of best fit line = 4.06, $p < 10^{-15}$).

Table 3: Distribution of p values for testing marker-lung cancer effect ($H_1 : \beta_G = 0|\beta_{GY} = 0$) and interaction between marker and smoking behavior on lung cancer risk ($H_2 : \beta_{GY=0}$), for SNPs identified as nominally significantly associated with smoking behavior ($p < 10^{-3}$) by the naïve or adjusted analysis. Each cell gives the number and percentage of SNPs whose p value for testing H_1 and H_2 fall within the respective range.

		$H_1 : \beta_G = 0\|\beta_{GY} = 0$			
		$[0.0, 0.1)$	$[0.1, 0.5)$	$[0.5, 1.0]$	Total
$H_2 : \beta_{GY} = 0$	$[0.0, 0.1)$	27	39	13	79
		3.3%	4.8%	1.6%	9.7%
	$[0.1, 0.5)$	104	134	94	332
		12.7%	16.4%	11.5%	40.6%
	$[0.5, 1.0]$	86	211	110	407
		10.5%	25.8%	13.4%	49.8%
	Total	217	384	217	818
		26.5%	46.9%	26.5%	100%

Table 4: Distribution of p values for testing marker-lung cancer effect ($H_1 : \beta_G = 0|\beta_{GY} = 0$) and interaction between marker and smoking behavior on lung cancer risk ($H_2 : \beta_{GY=0}$), for SNPs identified as nominally significantly associated with smoking behavior ($p < 10^{-3}$) by the control-only or IPW analysis. Each cell gives the number and percentage of SNPs whose p value for testing H_1 and H_2 fall within the respective range.

		$H_1 : \beta_G = 0 \mid \beta_{GY} = 0$			
		$[0.0, 0.1)$	$[0.1, 0.5)$	$[0.5, 1.0]$	Total
$H_2 : \beta_{GY} = 0$	$[0.0, 0.1)$	84	183	162	429
		12.9%	28.0%	24.8%	65.7%
	$[0.1, 0.5)$	60	74	56	190
		9.2%	11.3%	8.6%	29.1%
	$[0.5, 1.0]$	15	15	4	34
		2.3%	2.3%	0.6%	5.2%
	Total	159	272	222	653
		24.3%	41.7%	34.0%	100%

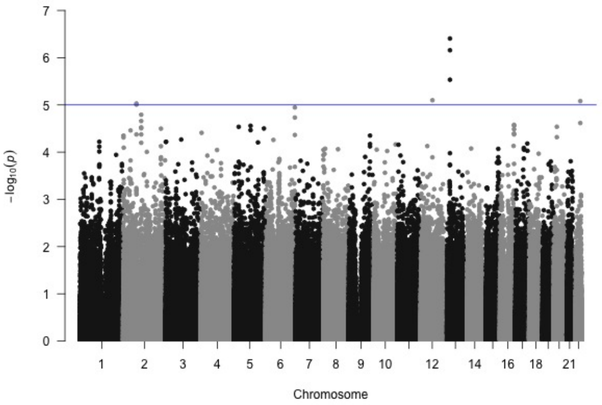

Figure 12: Manhattan plot for the naïve analysis of $\sqrt{\text{pack-years}}$. $-\log_{10} p$ values from a 1-DF Wald test for all SNPs passing quality control and assuming an additive genetic model. Analysis was performed using all 1,426 ever-smokers.

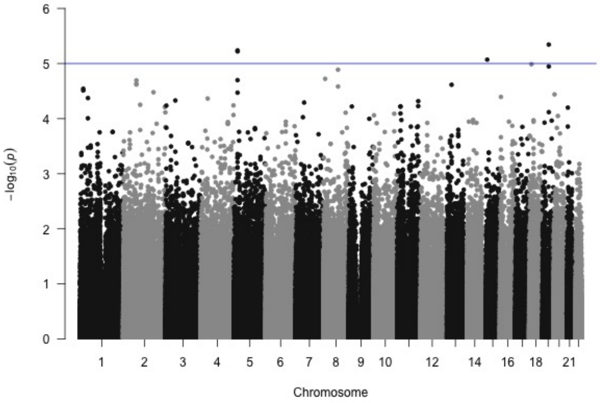

Figure 13: Manhattan plot for the control-only analysis of $\sqrt{\text{pack-years}}$. $-\log_{10} p$ values from a 1-DF Wald test for all SNPs passing quality control and assuming an additive genetic model. Analysis was performed using only the control ever-smokers ($n_0 = 730$).

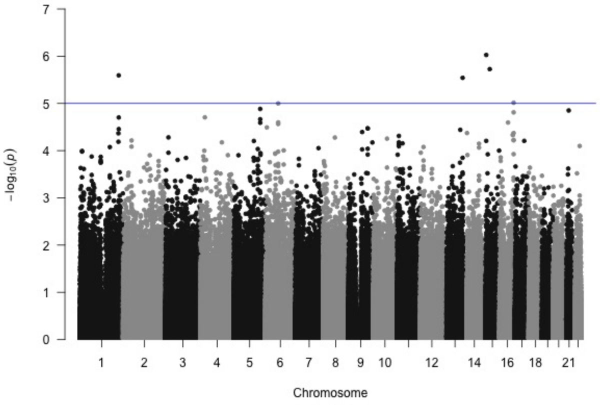

Figure 14: Manhattan plot for the case-only analysis of $\sqrt{\text{pack-years}}$. $-\log_{10} p$ values from a 1-DF Wald test for all SNPs passing quality control and assuming an additive genetic model. Analysis was performed using only the case ever-smokers ($n_1 = 696$).

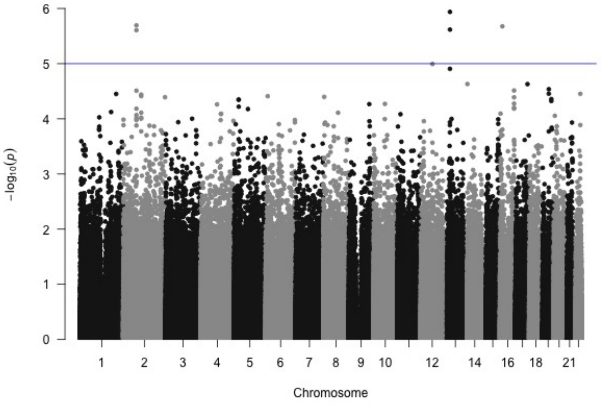

Figure 15: Manhattan plot for the adjusted analysis of $\sqrt{\text{pack-years}}$. $-\log_{10} p$ values from a 1-DF Wald test for all SNPs passing quality control and assuming an additive genetic model. Analysis was performed using all 1,426 ever-smokers.

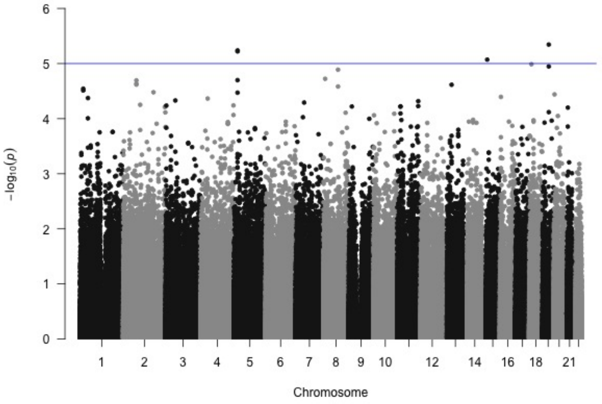

Figure 16: Manhattan plot for the IPW analysis of $\sqrt{\text{pack-years}}$. $-\log_{10} p$ values from a 1-DF Wald test for all SNPs passing quality control and assuming an additive genetic model. We estimate the prevalence of lung cancer among ever-smokers to be $\pi = 0.00148$ in order to calculate the inverse probability weights for all 1,426 study individuals.

Table 5: Top 10 SNPs from the genome-wide naïve analysis of $\sqrt{\text{pack-years}}$. Estimates of the additive genetic effect on smoking behavior ($\widehat{\alpha}_G$) and their p values from a 1-DF Wald test for the ad hoc methods and IPW regression. Marker-lung cancer effect estimates ($\widehat{OR}_{DG} = \exp(\widehat{\beta}_G)$), estimates for interaction between SNPs and smoking behavior on lung cancer risk ($\widehat{\beta}_{GY}$), and their p values from a 1-DF Wald test are also provided. Genes that have been identified in previous GWASs of smoking cessation are marked by asterisks.

			Lung cancer		**Smoking behavior**				
SNP	Chr.	Gene	$\widehat{OR}_{DG}$	$\widehat{\beta}_{GY}$	Naïve	Control-only	Case-only	Adjusted	IPW
rs3771823	2	*TACR1**	1.15	0.04	-0.40	-0.51	-0.30	-0.40	-0.51
			(1.13e-01)	(3.36e-01)	(9.35e-06)	(2.32e-05)	(1.13e-02)	(2.48e-06)	(1.02e-05)
rs7588326	2	*TACR1**	1.16	0.04	-0.40	-0.51	-0.30	-0.40	-0.51
			(8.41e-02)	(3.54e-01)	(9.82e-06)	(2.02e-05)	(9.97e-03)	(2.02e-06)	(8.82e-06)
rs11889631	2	*SLC9A2**	0.98	-0.02	0.51	0.48	0.43	0.45	0.48
			(8.42e-01)	(7.26e-01)	(1.61e-05)	(3.18e-03)	(4.15e-03)	(3.88e-05)	(3.29e-03)
rs7766185	6	*RPS6KA2**	0.94	0.05	-0.56	-0.51	-0.40	-0.46	-0.51
			(5.86e-01)	(3.87e-01)	(1.13e-05)	(1.67e-03)	(2.52e-02)	(1.05e-04)	(7.38e-04)
rs7771460	6	*RPS6KA2**	0.96	0.05	-0.55	-0.49	-0.43	-0.46	-0.49
			(7.23e-01)	(3.79e-01)	(1.85e-05)	(3.02e-03)	(1.42e-02)	(1.16e-04)	(1.55e-03)
rs10878841	12	N/A	1.07	0.02	-0.39	-0.38	-0.35	-0.36	-0.39
			(4.45e-01)	(6.45e-01)	(7.95e-06)	(9.68e-04)	(2.19e-03)	(1.01e-05)	(1.01e-03)
rs12172796	13	*NBEA**	0.96	-0.04	0.67	0.66	0.52	0.59	0.66
			(7.30e-01)	(4.85e-01)	(3.93e-07)	(2.76e-04)	(1.61e-03)	(1.15e-06)	(2.18e-04)
rs9788362	13	*NBEA**	0.97	-0.05	0.66	0.68	0.48	0.58	0.68
			(8.22e-01)	(3.99e-01)	(6.95e-07)	(2.49e-04)	(3.35e-03)	(2.41e-06)	(2.21e-04)
rs9574213	13	*NBEA**	1.00	-0.06	0.62	0.65	0.42	0.54	0.65
			(9.98e-01)	(2.87e-01)	(2.93e-06)	(4.23e-04)	(1.08e-02)	(1.24e-05)	(3.40e-04)
rs4823168	22	N/A	1.00	-0.00	-0.42	-0.36	-0.35	-0.37	-0.36
			(9.74e-01)	(9.41e-01)	(8.30e-06)	(3.98e-03)	(5.94e-03)	(3.53e-05)	(3.01e-03)

Table 6: Top 10 SNPs from the genome-wide control-only analysis of $\sqrt{\text{pack-years}}$. Estimates of the additive genetic effect on smoking behavior ($\widehat{\alpha}_G$) and their p values from a 1-DF Wald test for the ad hoc methods and IPW regression. Marker-lung cancer effect estimates ($\widehat{OR}_{DG} = \exp(\widehat{\beta}_G)$), estimates for interaction between SNPs and smoking behavior on lung cancer risk ($\widehat{\beta}_{GY}$), and their p values from a 1-DF Wald test are also provided. Genes that have been identified in previous GWASs of smoking cessation are marked by asterisks.

			Lung cancer		**Smoking behavior**				
SNP	Chr.	Gene	$\widehat{OR}_{DG}$	$\widehat{\beta}_{GY}$	Naïve	Control-only	Case-only	Adjusted	IPW
rs7588326	2	*TACR1**	1.16	0.04	-0.40	-0.51	-0.30	-0.40	-0.51
			(8.41e-02)	(3.54e-01)	(9.82e-06)	(2.02e-05)	(9.97e-03)	(2.02e-06)	(8.82e-06)
rs4461636	5	*CDH18**	1.25	0.24	-0.44	-0.98	0.01	-0.46	-0.98
			(1.44e-01)	(2.11e-03)	(5.57e-03)	(5.78e-06)	(9.49e-01)	(1.64e-03)	(7.78e-06)
rs4242066	5	*CDH18**	1.28	0.27	-0.40	-0.99	0.06	-0.44	-0.99
			(1.08e-01)	(8.24e-04)	(1.13e-02)	(6.02e-06)	(7.65e-01)	(2.93e-03)	(7.26e-06)
rs1391429	5	*CDH18**	1.22	0.24	-0.43	-0.90	-0.03	-0.45	-0.90
			(1.73e-01)	(1.75e-03)	(5.80e-03)	(2.00e-05)	(8.95e-01)	(1.99e-03)	(2.50e-05)
rs7842063	8	N/A	0.91	-0.10	0.44	0.64	0.17	0.41	0.64
			(3.79e-01)	(3.16e-02)	(8.69e-05)	(1.29e-05)	(2.52e-01)	(7.77e-05)	(4.41e-05)
rs4404875	8	*RP1L1*	0.86	-0.11	0.35	0.66	0.05	0.36	0.66
			(1.52e-01)	(2.06e-02)	(2.19e-03)	(1.89e-05)	(7.23e-01)	(6.89e-04)	(1.74e-05)
rs1655645	15	*FAM189A1**	0.95	-0.12	0.28	0.55	-0.03	0.26	0.55
			(5.89e-01)	(1.25e-03)	(1.99e-03)	(8.48e-06)	(8.05e-01)	(2.17e-03)	(2.08e-05)
rs1893213	18	N/A	0.89	-0.10	0.28	0.54	0.00	0.28	0.54
			(1.68e-01)	(1.40e-02)	(2.30e-03)	(1.03e-05)	(9.91e-01)	(1.04e-03)	(5.98e-06)
rs4805573	19	*ZNF536*	1.36	0.08	-0.65	-1.06	-0.27	-0.67	-1.06
			(6.63e-02)	(2.63e-01)	(1.77e-04)	(4.52e-06)	(2.38e-01)	(2.91e-05)	(1.29e-06)
rs4805574	19	*ZNF536*	1.34	0.07	-0.65	-1.02	-0.30	-0.67	-1.02
			(8.39e-02)	(3.43e-01)	(1.84e-04)	(1.13e-05)	(1.85e-01)	(3.51e-05)	(3.99e-06)

Table 7: Top 10 SNPs from the genome-wide case-only analysis of $\sqrt{\text{pack-years}}$. Estimates of the additive genetic effect on smoking behavior ($\widehat{\alpha}_G$) and their p values from a 1-DF Wald test for the ad hoc methods and IPW regression. Marker-lung cancer effect estimates ($\widehat{OR}_{DG} = \exp(\widehat{\beta}_G)$), estimates for interaction between SNPs and smoking behavior on lung cancer risk ($\widehat{\beta}_{GY}$), and their p values from a 1-DF Wald test are also provided. Genes that have been identified in previous GWASs of smoking cessation are marked by asterisks.

			Lung cancer		Smoking behavior				
SNP	Chr.	Gene	$\widehat{OR}_{DG}$	$\widehat{\beta}_{GY}$	Naïve	Control-only	Case-only	Adjusted	IPW
rs1108089	1	N/A	1.27	-0.18	-0.45	-0.01	-0.92	-0.48	-0.01
			(9.95e-02)	(4.25e-03)	(4.38e-03)	(9.62e-01)	(2.55e-06)	(9.70e-04)	(9.58e-01)
rs959903	4	SEL1L3*	1.02	-0.12	-0.26	0.05	-0.56	-0.24	0.05
			(8.26e-01)	(4.78e-03)	(8.29e-03)	(6.84e-01)	(1.97e-05)	(1.09e-02)	(6.76e-01)
rs415426	5	SLC36A1*	1.02	0.13	0.20	-0.15	0.48	0.16	-0.15
			(8.26e-01)	(9.38e-04)	(2.47e-02)	(2.03e-01)	(1.31e-05)	(4.78e-02)	(2.00e-01)
rs9341360	6	RIMS1*	0.84	0.08	0.25	0.06	0.52	0.28	0.06
			(4.07e-02)	(3.99e-02)	(5.77e-03)	(6.05e-01)	(1.00e-05)	(8.60e-04)	(5.86e-01)
rs7317390	13	COL4A2	1.21	-0.03	-0.25	0.02	-0.61	-0.29	0.02
			(4.33e-02)	(5.64e-01)	(1.31e-02)	(8.78e-01)	(2.87e-06)	(2.63e-03)	(8.73e-01)
rs4906879	15	LOC105370740	0.78	0.14	0.12	-0.13	0.58	0.20	-0.13
			(4.59e-03)	(8.19e-04)	(1.75e-01)	(2.98e-01)	(9.41e-07)	(1.94e-02)	(2.96e-01)
rs1484197	15	N/A	0.87	0.09	0.32	0.06	0.60	0.32	0.06
			(1.55e-01)	(4.47e-02)	(1.43e-03)	(6.42e-01)	(1.88e-06)	(4.92e-04)	(6.32e-01)
rs2966249	16	HSD17B2*	0.96	0.06	0.38	0.17	0.52	0.35	0.17
			(6.72e-01)	(1.63e-01)	(4.04e-05)	(1.75e-01)	(9.68e-06)	(6.31e-05)	(1.77e-01)
rs1017243	16	HSD17B2*	0.96	0.05	0.38	0.18	0.51	0.34	0.18
			(6.74e-01)	(1.81e-01)	(4.22e-05)	(1.53e-01)	(1.55e-05)	(6.54e-05)	(1.55e-01)
rs2829949	21	N/A	0.95	0.22	0.54	-0.08	1.03	0.48	-0.08
			(7.56e-01)	(3.70e-02)	(4.33e-03)	(7.59e-01)	(1.41e-05)	(6.31e-03)	(7.45e-01)

Table 8: Top 10 SNPs from the genome-wide adjusted analysis of $\sqrt{\text{pack-years}}$. Estimates of the additive genetic effect on smoking behavior ($\widehat{\alpha}_G$) and their p values from a 1-DF Wald test for the ad hoc methods and IPW regression. Marker-lung cancer effect estimates ($\widehat{OR}_{DG} = \exp(\widehat{\beta}_G)$), estimates for interaction between SNPs and smoking behavior on lung cancer risk ($\widehat{\beta}_{GY}$), and their p values from a 1-DF Wald test are also provided. Genes that have been identified in previous GWASs of smoking cessation are marked by asterisks.

			Lung cancer		Smoking behavior				
SNP	Chr.	Gene	$\widehat{OR}_{DG}$	$\widehat{\beta}_{GY}$	Naïve	Control-only	Case-only	Adjusted	IPW
rs7588326	2	TACR1*	1.16	0.04	-0.40	-0.51	-0.30	-0.40	-0.51
			(8.41e-02)	(3.54e-01)	(9.82e-06)	(2.02e-05)	(9.97e-03	(2.02e-06)	(8.82e-06)
rs3771823	2	TACR1*	1.15	0.04	-0.40	-0.51	-0.30	-0.40	-0.51
			(1.13e-01)	(3.36e-01)	(9.35e-06)	(2.32e-05)	(1.13e-02)	(2.48e-06)	(1.02e-05)
rs10878841	12	N/A	1.07	0.02	-0.39	-0.38	-0.35	-0.36	-0.39
			(4.45e-01)	(6.45e-01)	(7.95e-06)	(9.68e-04)	(2.19e-03)	(1.01e-05)	(1.01e-03)
rs12172796	13	NBEA*	0.96	-0.04	0.67	0.66	0.52	0.59	0.66
			(7.30e-01)	(4.85e-01)	(3.93e-07	2.76e-04)	(1.61e-03)	(1.15e-06)	(2.18e-04)
rs9788362	13	NBEA*	0.97	-0.05	0.66	0.68	0.48	0.58	0.68
			(8.22e-01)	(3.99e-01)	(6.95e-07	2.49e-04)	(3.35e-03)	(2.41e-06)	(2.21e-04)
rs9574213	13	NBEA*	1.00	-0.06	0.62	0.65	0.42	0.54	0.65
			(9.98e-01)	(2.87e-01)	(2.93e-06)	(4.23e-04)	(1.08e-02)	(1.24e-05)	(3.40e-04)
rs1952512	14	TMEM253	0.72	0.09	0.59	0.42	0.87	0.64	0.42
			(3.12e-02)	(2.25e-01)	(2.69e-04)	(5.27e-02)	(4.24e-05)	(2.34e-05)	(4.94e-02)
rs8053423	16	N/A	0.69	0.01	0.39	0.48	0.44	0.47	0.48
			(2.55e-04)	(8.69e-01)	(3.07e-04)	(5.25e-04)	(2.34e-03)	(2.11e-06)	(2.25e-04)
rs12941222	17	SDK2	0.87	-0.03	0.38	0.48	0.29	0.38	0.48
			(1.39e-01)	(4.32e-01)	(9.36e-05)	(2.30e-04)	(2.10e-02)	(2.35e-05)	(1.28e-04)
rs4805573	19	ZNF536	1.36	0.08	-0.65	-1.06	-0.27	-0.67	-1.06
			(6.63e-02)	(2.63e-01)	(1.77e-04)	(4.52e-06)	(2.38e-01)	(2.91e-05)	(1.29e-06)

Table 9: Top 10 SNPs from the genome-wide IPW analysis of $\sqrt{\text{pack-years}}$. Estimates of the additive genetic effect on smoking behavior ($\widehat{\alpha}_G$) and their p values from a 1-DF Wald test for the ad hoc methods and IPW regression. Marker-lung cancer effect estimates ($\widehat{OR}_{DG} = \exp(\widehat{\beta}_G)$), estimates for interaction between SNPs and smoking behavior on lung cancer risk ($\widehat{\beta}_{GY}$), and their p values from a 1-DF Wald test are also provided. Genes that have been identified in previous GWASs of smoking cessation are marked by asterisks.

			Lung cancer		**Smoking behavior**				
SNP	Chr.	Gene	$\widehat{OR}_{DG}$	$\widehat{\beta}_{GY}$	Naïve	Control-only	Case-only	Adjusted	IPW
rs7588326	2	*TACR1**	1.16	0.04	-0.40	-0.51	-0.30	-0.40	-0.51
			(8.41e-02)	(3.54e-01)	(9.82e-06)	(2.02e-05)	(9.97e-03)	(2.02e-06)	(8.82e-06)
rs3771823	2	*TACR1**	1.15	0.04	-0.40	-0.51	-0.30	-0.40	-0.51
			(1.13e-01)	(3.36e-01)	(9.35e-06)	(2.32e-05)	(1.13e-02)	(2.48e-06)	(1.02e-05)
rs741418	2	*TACR1**	1.08	0.04	-0.38	-0.51	-0.19	-0.36	-0.51
			(3.68e-01)	(2.74e-01)	(4.22e-05)	(2.41e-05)	(1.13e-01)	(3.09e-05)	(1.19e-05)
rs4242066	5	*CDH18**	1.28	0.27	-0.40	-0.99	0.06	-0.44	-0.99
			(1.08e-01)	(8.24e-04)	(1.13e-02)	(6.02e-06)	(7.65e-01)	(2.93e-03)	(7.26e-06)
rs4461636	5	*CDH18**	1.25	0.24	-0.44	-0.98	0.01	-0.46	-0.98
			(1.44e-01)	(2.11e-03)	(5.57e-03)	(5.78e-06)	(9.49e-01)	(1.64e-03)	(7.78e-06)
rs4942376	13	N/A	0.87	-0.07	0.32	0.58	0.06	0.33	0.58
			(1.63e-01)	(1.40e-01)	(1.74e-03)	(2.43e-05)	(6.72e-01)	(6.47e-04)	(7.89e-06)
rs1893213	18	N/A	0.89	-0.10	0.28	0.54	0.00	0.28	0.54
			(1.68e-01)	(1.40e-02)	(2.30e-03)	(1.03e-05)	(9.91e-01)	(1.04e-03)	(5.98e-06)
rs4805573	19	*ZNF536*	1.36	0.08	-0.65	-1.06	-0.27	-0.67	-1.06
			(6.63e-02)	(2.63e-01)	(1.77e-04)	(4.52e-06)	(2.38e-01)	(2.91e-05)	(1.29e-06)
rs4805574	19	*ZNF536*	1.34	0.07	-0.65	-1.02	-0.30	-0.67	-1.02
			(8.39e-02)	(3.43e-01)	(1.84e-04)	(1.13e-05)	(1.85e-01)	(3.51e-05)	(3.99e-06)
rs6052961	20	*SLC23A2*	1.08	0.12	-0.26	-0.52	0.03	-0.25	-0.52
			(3.50e-01)	(4.20e-03)	(4.72e-03)	(3.63e-05)	(8.17e-01)	(2.84e-03)	(7.46e-06)

Table 10: Top 10 novel SNPs from the genome-wide control-only analysis of $\sqrt{\text{pack-years}}$. Estimates of the additive genetic effect on smoking behavior ($\widehat{\alpha}_G$) and their p values from a 1-DF Wald test for the ad hoc methods and IPW regression. SNPs are novel in the sense that they are nominally significant ($p < 10^{-3}$) when analyzed by the control-only analysis, but nominally insignificant ($p \geq 10^{-3}$) when analyzed by the adjusted and IPW analyses. Marker-lung cancer effect estimates ($\widehat{OR}_{DG} = \exp(\widehat{\beta}_G)$), estimates for interaction between SNPs and smoking behavior on lung cancer risk ($\widehat{\beta}_{GY}$), and their p values from a 1-DF Wald test are also provided. Genes that have been identified in previous GWASs of smoking cessation are marked by asterisks.

			Lung cancer		**Smoking behavior**				
SNP	Chr.	Gene	$\widehat{OR}_{DG}$	$\widehat{\beta}_{GY}$	Naïve	Control-only	Case-only	Adjusted	IPW
rs1568340	1	N/A	0.88	-0.12	0.23	0.49	-0.02	0.25	0.49
			(1.94e-01)	(3.47e-03)	(3.56e-02)	(7.05e-04)	(9.11e-01)	(1.35e-02)	(1.48e-03)
rs7643114	3	N/A	0.79	-0.20	0.19	0.60	-0.13	0.25	0.60
			(4.52e-02)	(1.03e-05)	(1.19e-01)	(2.06e-04)	(4.17e-01)	(2.86e-02)	(1.25e-03)
rs4955322	3	N/A	0.76	-0.20	0.19	0.58	-0.09	0.26	0.58
			(2.06e-02)	(2.34e-05)	(1.25e-01)	(3.20e-04)	(5.65e-01)	(2.23e-02)	(1.71e-03)
rs950206	4	N/A	1.07	0.02	-0.25	-0.43	-0.05	-0.24	-0.42
			(4.48e-01)	(6.06e-01)	(7.02e-03)	(7.53e-04)	(6.86e-01)	(4.93e-03)	(1.05e-03)
rs221723	6	PDE10A*	0.88	-0.09	0.24	0.45	0.02	0.25	0.44
			(1.58e-01)	(2.40e-02)	(1.32e-02)	(4.68e-04)	(8.41e-01)	(4.89e-03)	(1.03e-03)
rs221725	6	PDE10A*	0.87	-0.09	0.23	0.44	0.02	0.25	0.44
			(1.36e-01)	(2.87e-02)	(1.57e-02)	(5.45e-04)	(8.41e-01)	(5.30e-03)	(1.22e-03)
rs596077	11	LOC105369591	1.07	-0.14	0.19	0.45	-0.18	0.14	0.45
			(4.63e-01)	(2.52e-04)	(5.67e-02)	(5.94e-04)	(1.43e-01)	(1.15e-01)	(1.16e-03)
rs1241482	14	LOC105370414	0.96	-0.13	0.29	0.51	-0.01	0.26	0.51
			(7.35e-01)	(3.24e-03)	(1.09e-02)	(6.61e-04)	(9.34e-01)	(1.19e-02)	(1.35e-03)
rs2289567	17	KSR1*	0.97	-0.17	0.19	0.53	-0.16	0.17	0.53
			(7.92e-01)	(2.55e-04)	(1.02e-01)	(6.84e-04)	(2.68e-01)	(1.07e-01)	(1.50e-03)
rs10513972	18	DOK6*	1.30	0.13	0.01	-0.58	0.36	-0.08	-0.58
			(2.59e-02)	(2.20e-02)	(9.04e-01)	(7.35e-04)	(1.86e-02)	(5.06e-01)	(1.05e-03)

Table 11: Top 10 novel SNPs from the genome-wide adjusted analysis of $\sqrt{\text{pack-years}}$. Estimates of the additive genetic effect on smoking behavior ($\widehat{\alpha}_G$) and their p values from a 1-DF Wald test for the ad hoc methods and IPW regression. SNPs are novel in the sense that they are nominally significant ($p < 10^{-3}$) when analyzed by the adjusted analysis, but nominally insignificant ($p \geq 10^{-3}$) when analyzed by the control-only and IPW analyses. Marker-lung cancer effect estimates ($\widehat{OR}_{DG} = \exp(\widehat{\beta}_G)$), estimates for interaction between SNPs and smoking behavior on lung cancer risk ($\widehat{\beta}_{GY}$), and their p values from a 1-DF Wald test are also provided. Genes that have been identified in previous GWASs of smoking cessation are marked by asterisks.

			Lung cancer		**Smoking behavior**				
SNP	Chr.	Gene	$\widehat{OR}_{DG}$	$\widehat{\beta}_{GY}$	Naïve	Control-only	Case-only	Adjusted	IPW
rs696981	1	N/A	0.90	0.01	0.35	-0.40	0.28	0.35	-0.40
			(1.94e-01)	(8.65e-01)	(1.14e-04)	(1.09e-03)	(1.71e-02)	(3.55e-05)	(1.16e-03)
rs4851022	2	SLC9A2*	0.96	-0.01	0.50	0.48	0.43	0.45	0.48
			(6.82e-01)	(8.67e-01)	(2.20e-05)	(3.24e-03)	(4.06e-03)	(3.64e-05)	(3.13e-03)
rs11889631	2	SLC9A2*	0.98	-0.02	0.51	0.48	0.43	0.45	0.48
			(8.42e-01)	(7.26e-01)	(1.61e-05)	(3.18e-03)	(4.15e-03)	(3.88e-05)	(3.29e-03)
rs6431588	2	ILKAP	1.21	0.04	-0.47	-0.47	-0.47	-0.47	-0.47
			(1.01e-01)	(4.23e-01)	(1.49e-04)	(6.15e-03)	(2.09e-03)	(4.06e-05)	(4.34e-03)
rs4540426	8	N/A	0.77	0.04	0.51	0.41	0.68	0.54	0.41
			(4.50e-02)	(4.62e-01)	(2.68e-04)	(3.01e-02)	(1.67e-04)	(4.00e-05)	(3.45e-02)
rs1952512	14	TMEM253	0.72	0.09	0.59	0.42	0.87	0.64	0.42
			(3.12e-02)	(2.25e-01)	(2.69e-04)	(5.27e-02)	(4.24e-05)	(2.34e-05)	(4.94e-02)
rs10514525	16	N/A	0.95	0.04	0.38	0.24	0.47	0.35	0.24
			(5.13e-01)	(3.16e-01)	(2.75e-05)	(5.44e-02)	(4.24e-05)	(3.06e-05)	(4.82e-02)
rs4888202	16	HSD17B2*	0.96	0.04	0.38	0.23	0.47	0.35	0.23
			(6.62e-01)	(3.46e-01)	(2.64e-05)	(6.67e-02)	(5.94e-05)	(4.08e-05)	(6.66e-02)
rs8111069	19	CLPTM1	1.18	-0.04	-0.35	-0.34	-0.38	-0.37	-0.34
			(6.58e-02)	(3.93e-01)	(2.43e-04)	(9.62e-03)	(2.15e-03)	(4.39e-05)	(6.16e-03)
rs4823168	22	N/A	1.00	-0.00	-0.42	-0.36	-0.35	-0.37	-0.36
			(9.74e-01)	(9.41e-01)	(8.30e-06)	(3.98e-03)	(5.94e-03)	(3.53e-05)	(3.01e-03)

Table 12: Top 10 novel SNPs from the genome-wide IPW analysis of $\sqrt{\text{pack-years}}$. Estimates of the additive genetic effect on smoking behavior ($\widehat{\alpha}_G$) and their p values from a 1-DF Wald test for the ad hoc methods and IPW regression. SNPs are novel in the sense that they are nominally significant ($p < 10^{-3}$) when analyzed by IPW regression, but nominally insignificant ($p \geq 10^{-3}$) when analyzed by the control-only and adjusted analyses. Marker-lung cancer effect estimates ($\widehat{OR}_{DG} = \exp(\widehat{\beta}_G)$), estimates for interaction between SNPs and smoking behavior on lung cancer risk ($\widehat{\beta}_{GY}$), and their p values from a 1-DF Wald test are also provided. Genes that have been identified in previous GWASs of smoking cessation are marked by asterisks.

			Lung cancer		**Smoking behavior**				
SNP	Chr.	Gene	$\widehat{OR}_{DG}$	$\widehat{\beta}_{GY}$	Naïve	Control-only	Case-only	Adjusted	IPW
rs6691873	1	*C1orf95**	1.18	0.34	-0.29	-0.82	0.21	-0.30	-0.82
			(3.53e-01)	(2.77e-03)	(1.31e-01)	(1.47e-03)	(3.81e-01)	(8.49e-02)	(2.11e-04)
rs1552290	1	N/A	1.24	-0.16	0.28	0.65	-0.21	0.17	0.65
			(9.96e-02)	(3.24e-03)	(4.36e-02)	(1.18e-03)	(2.17e-01)	(1.99e-01)	(3.22e-04)
rs11722134	4	N/A	1.23	-0.07	0.30	0.71	-0.24	0.19	0.71
			(1.73e-01)	(3.14e-01)	(6.10e-02)	(1.92e-03)	(2.34e-01)	(2.00e-01)	(1.74e-04)
rs1316405	4	*ARHGAP24**	1.04	0.26	-0.33	-0.76	0.18	-0.31	-0.75
			(8.00e-01)	(1.11e-02)	(6.69e-02)	(1.43e-03)	(4.58e-01)	(6.72e-02)	(2.92e-04)
rs1870658	5	*LOC105379037*	1.29	0.23	-0.40	-0.72	-0.15	-0.44	-0.72
			(9.81e-02)	(4.96e-03)	(1.34e-02)	(1.02e-03)	(4.63e-01)	(3.25e-03)	(6.45e-05)
rs471405	5	*LOC105379037*	1.29	0.23	-0.40	-0.72	-0.15	-0.44	-0.72
			(9.81e-02)	(4.96e-03)	(1.34e-02)	(1.02e-03)	(4.63e-01)	(3.25e-03)	(6.45e-05)
rs26008	5	*FNIP1*	1.11	0.27	-0.56	-0.85	-0.17	-0.53	-0.84
			(5.68e-01)	(1.39e-02)	(4.58e-03)	(1.07e-03)	(5.28e-01)	(4.04e-03)	(2.58e-04)
rs2642576	10	N/A	1.05	-0.19	0.10	0.69	-0.58	0.07	0.69
			(7.27e-01)	(6.02e-03)	(5.47e-01)	(2.09e-03)	(6.64e-03)	(6.47e-01)	(2.20e-04)
rs9921063	16	*CDYL2**	0.95	0.32	-0.35	-0.67	0.14	-0.29	-0.67
			(7.39e-01)	(3.77e-04)	(2.72e-02)	(1.09e-03)	(4.89e-01)	(4.82e-02)	(2.29e-04)
rs549300	18	N/A	1.10	0.14	-0.35	-0.63	-0.06	-0.34	-0.63
			(4.70e-01)	(3.93e-02)	(1.53e-02)	(1.08e-03)	(7.37e-01)	(1.14e-02)	(3.82e-04)