SUPPLEMENTARY DATA

Supplementary methods

Mendelian randomization framework

Let $\hat{\Gamma}_j$ equal the gene-outcome association estimate for variant j = 1, . . ., J, with associated standard error σ_{Yj}. Let $\hat{\gamma}_j$ equal the gene-exposure association estimate for variant j, with associated standard error σ_{Xj}. Let the causal effect of the exposure on the outcome be denoted by β. An estimate for β based on variant j alone can be obtained via the ratio method as

$$\hat{\beta}_j = \frac{\hat{\Gamma}_j}{\hat{\gamma}_j}$$

Two forms for the variance of $\hat{\beta}_j$ are often used:

(i) $Var(\hat{\beta}_j) = \dfrac{\sigma_{Yj}^2}{\hat{\gamma}_j^2}$

(ii) $Var(\hat{\beta}_j) = \dfrac{\sigma_{Yj}^2}{\hat{\gamma}_j^2} + \dfrac{\hat{\Gamma}_j^2 \sigma_{Xj}^2}{\hat{\gamma}_j^4}$,

Using either a first order (i) or second order (ii) Taylor series expansion. We use the variance from (i). This is equivalent to assuming that the gene-exposure association estimates are measured without error and is referred to as the No Measurement Error (NOME) assumption. NOME is equivalent to the assumption $\sigma_{X\,j}^2 = 0$ for all j, so that $\hat{\gamma}_j = \gamma_j$ for all j.

The inverse variance weighted (IVW) method for the overall causal effect estimate

Let $w_j = 1/\operatorname{var}(\hat{\beta}_j)$ where $\operatorname{var}(\hat{\beta}_j)$is defined as in (i) under NOME. The inverse variance weighted (IVW) estimate for the causal effect is given by the standard meta-analytic formula

$$\frac{\sum_j w_j \hat{\beta}_j}{\sum_j w_j}.$$

The w_j terms derived under NOME are also referred to as 'Toby Johnson' weights. The IVW estimate assumes that all genetic variants satisfy the instrumental variable assumptions. If this is not true then it could give a biased estimate for β.The IVW estimate for β is consistent even if all genetic variants are invalid, provided that:

- Across all variants, the magnitude of the gene exposure associations are independent of their pleiotropic effects (the InSIDE assumption)
- NOME is satisfied
- The pleiotropic effects have zero mean

The weighted median method for the overall causal effect estimate

Let $\hat{\beta}_{(1)},\ldots,\hat{\beta}_{(J)}$ equal the J causal effect estimates ordered from smallest ($\hat{\beta}_{(1)}$) to largest ($\hat{\beta}_{(J)}$). Now define

$$w^{*}_{(j)} = \frac{w_j}{S_J}, \quad \text{where} \quad S_J = \sum_j w_j,$$

and equate $\hat{\beta}_{(j)}$with a quantile, $p^{w}_{(j)}$, defined as

$$p^{w}_{(j)} = \frac{100}{S_J}\left(S_{(j)} - \frac{w_{(j)}}{2}\right).$$

$p^{w}_{(j)}$ represents the quantile from the weighted empirical distribution function of the ordered estimates$\hat{\beta}_{(1)},\ldots,\hat{\beta}_{(J)}$. The weighted median estimate, $\hat{\beta}_{WM}$ is defined as the 50th percentile of this weighted distribution. Typically the 50th percentile will lie between two estimates ($\hat{\beta}_{(l)}$and $\hat{\beta}_{(m)}$, say), in which case $\hat{\beta}_{WM}$ is found by linear interpolation.

$\hat{\beta}_{WM}$ is a consistent estimate for βprovided that at least 50% of the `weight' making up S_J comes from genetic variants that are valid instruments.

The MR-Egger method for the overall causal effect estimate

The MR-Egger method performs a weighted linear regression of the gene-outcome coefficients on the gene-exposure coefficients:

$$\frac{\hat{\Gamma}_j}{\sigma_{Yj}} = \frac{\beta_{0E}}{\sigma_{Yj}} + \beta_{1E}\frac{\hat{\gamma}_j}{\sigma_{Yj}}$$

The weights used are also derived under the NOME assumption. If all genetic variants are valid instruments, then β_{0E}= 0. The value of $\hat{\beta}_{0E}$ can be interpreted as an estimate of the average pleiotropic effect across the genetic variants. An intercept term that differs from zero is indicative of overall directional pleiotropy. The MR-Egger estimate for β, $\hat{\beta}_{1E}$, is consistent even if all genetic variants are invalid, provided that:

- Across all variants, the magnitude of the gene exposure associations are independent of their pleiotropic effects (the InSIDE assumption)
- NOME is satisfied.

If NOME is violated then the MR-Egger estimate of causal effect will be attenuated towards the null. We can assess the strength of NOME violation for MR-Egger through the I^2_{GX} statistic: $I^2_{GX} = \frac{Q - df}{Q}$, where $Q = \sum_{J=1}^{J} \frac{\left(\hat{\gamma}_j / \sigma^2_{Y_j} - \overline{\gamma}\right)^2}{\sigma^2_{X_j} / \sigma^2_{Y_j}}$ and where $\overline{\gamma}$ equals the arithmetic mean of the $\hat{\gamma}_j / \sigma^2_{Y\,j}$ terms . Specifically, the I^2_{GX} statistic quantifies the proportion of the total variation between the $\hat{\gamma}_j / \sigma^2_{Y\,j}$ terms that is due to `true' variation between the $\hat{\gamma}_j / \sigma^2_{Y\,j}$ terms. Consequently, when NOME is satisfied $\hat{\gamma}_1,\ldots,\hat{\gamma}_J = \gamma_1,\ldots,\gamma_J$, I^2_{GX} equals 1, and no attenuation occurs. When I^2_{GX} = 0.9 we can expect the MR-Egger estimate to be only 90% of its value had NOME been satisfied. A crude correction for NOME violation would be $\frac{\hat{\beta}_1 E}{I^2_{GX}}$, however this can be unstable as I^2_{GX} can sometimes be estimated as zero, even when it is truly large. We used the established method of Simulation Extrapolation (SIMEX) (1) instead, as implemented using the R package simex() (2). Under SIMEX, new data sets are created by simulating gene-exposure association estimates under increasing violations of NOME and recording the amount of attenuation in the estimate that occurs. The set of attenuated estimates are then used to extrapolate back to the estimate that would have been obtained if NOME had been satisfied.

SUPPLEMENTARY DATA

Supplementary Results

Outlier analysis – Studentized residuals

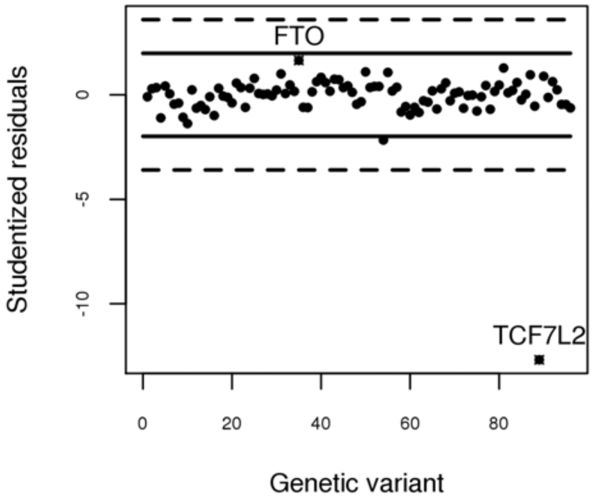

Figure S1A – Studentised residuals applied to the IVW method.

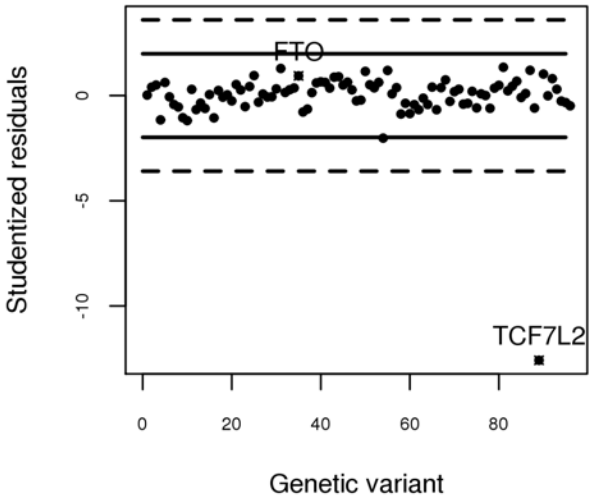

Figure S1B – Studentised residuals applied to the MR-Egger method.s

SUPPLEMENTARY DATA

Outlier analysis – Cook's distance

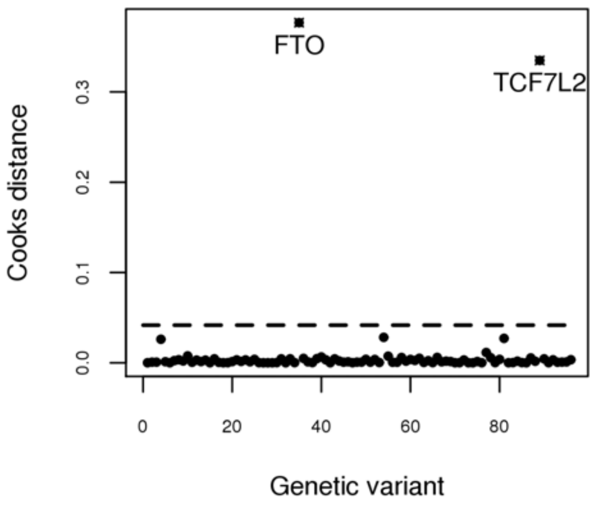

Figure S2A – Cook's distance applied to the IVW method.

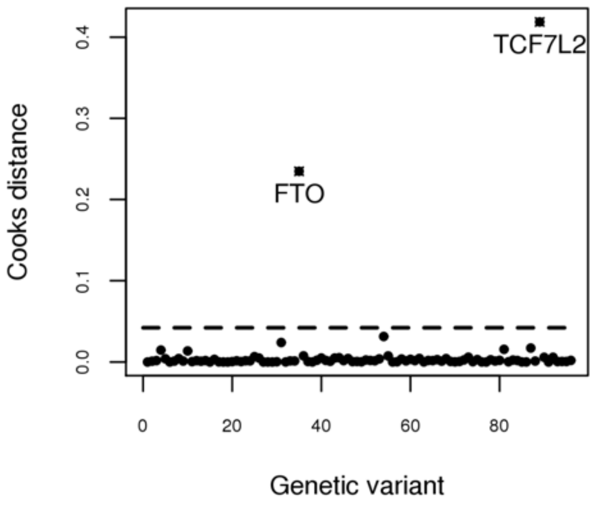

Figure S2B – Cook's distance applied to the MR-Egger method.

SUPPLEMENTARY DATA

Reciprocal analysis of type 2 diabetes and BMI

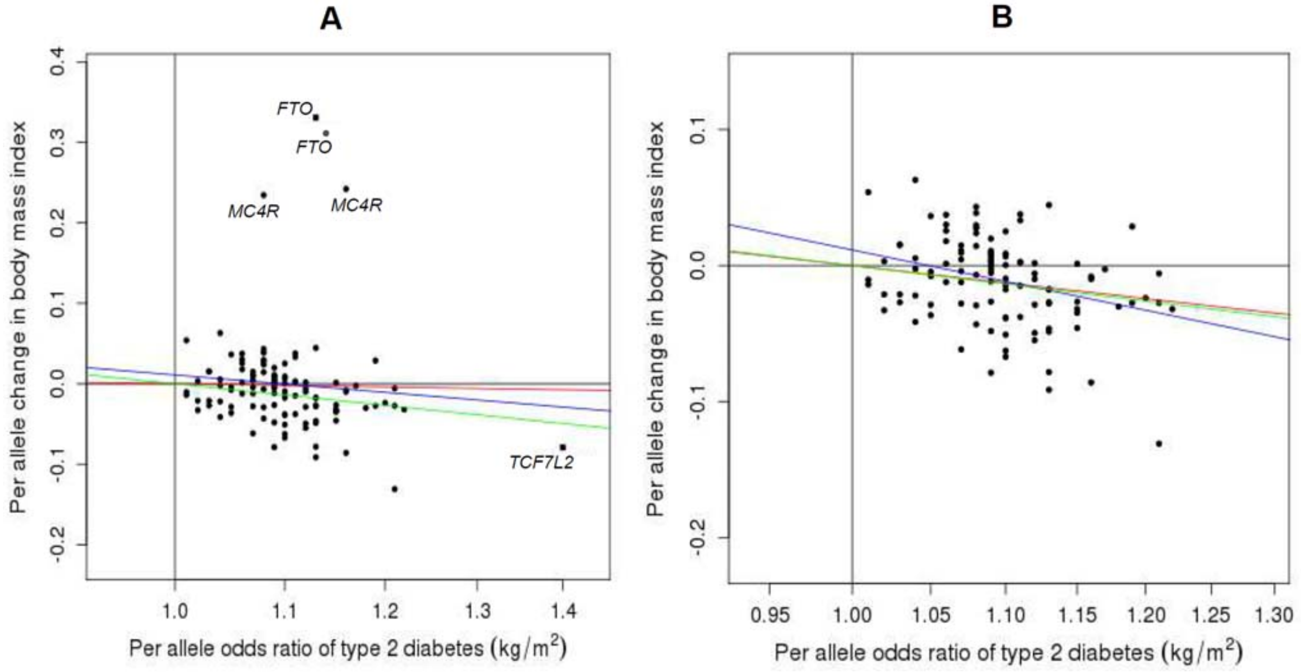

Figure S3 – MR-Egger analysis of the causal impact of type 2 diabetes on BMI.

A - scatter plot of genetic associations with BMI against associations with type 2 diabetes, with causal estimates (β coefficients) of type 2 diabetes on BMI estimated by inverse-variance weighted (red line), MR-Egger (blue line) and median-based (green line) methods. For this analysis, all 115 confirmed type 2 diabetes associated loci with OR not equal to 1 from Morris et al (2012)(3) downloaded from DIAGRAM http://diagram-consortium.org/downloads.html) were used.

A - scatter plot of genetic associations with BMI against associations with type 2 diabetes, with causal estimates (β coefficients) of type 2 diabetes on BMI estimated by inverse-variance weighted (red line), MR-Egger (blue line) and median-based (green line) methods. For this analysis, 110 confirmed type 2 diabetes associated loci with OR not equal to 1 and no overlapping known BMI loci (excluding *FTO*, *MC4R* and *TCF7L2*) from Morris et al (2012)(3)were again used.

SUPPLEMENTARY DATA

References

1. Cook JR, Stefanski LA: Simulation-Extrapolation Estimation in Parametric Measurement Error Models. Journal of the American Statistical Association 1994;89:1314-1328

2. Lederer W, Küchenhoff H: simex: SIMEX- and MCSIMEX-Algorithm for measurement error models., 2013

3. Morris AP, Voight BF, Teslovich TM, Ferreira T, Segrè AV, Steinthorsdottir V, Strawbridge RJ, Khan H, Grallert H, Mahajan A, Prokopenko I, Kang HM, Dina C, Esko T, Fraser RM, Kanoni S, Kumar A, Lagou V, Langenberg C, Luan Ja, Lindgren CM, Müller-Nurasyid M, Pechlivanis S, Rayner NW, Scott LJ, Wiltshire S, Yengo L, Kinnunen L, Rossin EJ, Raychaudhuri S, Johnson AD, Dimas AS, Loos RJF, Vedantam S, Chen H, Florez JC, Fox C, Liu C-T, Rybin D, Couper DJ, Kao WHL, Li M, Cornelis MC, Kraft P, Sun Q, van Dam RM, Stringham HM, Chines PS, Fischer K, Fontanillas P, Holmen OL, Hunt SE, Jackson AU, Kong A, Lawrence R, Meyer J, Perry JRB, Platou CGP, Potter S, Rehnberg E, Robertson N, Sivapalaratnam S, Stančáková A, Stirrups K, Thorleifsson G, Tikkanen E, Wood AR, Almgren P, Atalay M, Benediktsson R, Bonnycastle LL, Burtt N, Carey J, Charpentier G, Crenshaw AT, Doney ASF, Dorkhan M, Edkins S, Emilsson V, Eury E, Forsen T, Gertow K, Gigante B, Grant GB, Groves CJ, Guiducci C, Herder C, Hreidarsson AB, Hui J, James A, Jonsson A, Rathmann W, Klopp N, Kravic J, Krjutškov K, Langford C, Leander K, Lindholm E, Lobbens S, Männistö S, Mirza G, Mühleisen TW, Musk B, Parkin M, Rallidis L, Saramies J, Sennblad B, Shah S, Sigurðsson G, Silveira A, Steinbach G, Thorand B, Trakalo J, Veglia F, Wennauer R, Winckler W, Zabaneh D, Campbell H, van Duijn C, Uitterlinden AG, Hofman A, Sijbrands E, Abecasis GR, Owen KR, Zeggini E, Trip MD, Forouhi NG, Syvänen A-C, Eriksson JG, Peltonen L, Nöthen MM, Balkau B, Palmer CNA, Lyssenko V, Tuomi T, Isomaa B, Hunter DJ, Qi L, Wellcome Trust Case Control C, Investigators M, Consortium G, Consortium A-TD, Consortium SD, Shuldiner AR, Roden M, Barroso I, Wilsgaard T, Beilby J, Hovingh K, Price JF, Wilson JF, Rauramaa R, Lakka TA, Lind L, Dedoussis G, Njølstad I, Pedersen NL, Khaw K-T, Wareham NJ, Keinanen-Kiukaanniemi SM, Saaristo TE, Korpi-Hyövälti E, Saltevo J, Laakso M, Kuusisto J, Metspalu A, Collins FS, Mohlke KL, Bergman RN, Tuomilehto J, Boehm BO, Gieger C, Hveem K, Cauchi S, Froguel P, Baldassarre D, Tremoli E, Humphries SE, Saleheen D, Danesh J, Ingelsson E, Ripatti S, Salomaa V, Erbel R, Jöckel K-H, Moebus S, Peters A, Illig T, de Faire U, Hamsten A, Morris AD, Donnelly PJ, Frayling TM, Hattersley AT, Boerwinkle E, Melander O, Kathiresan S, Nilsson PM, Deloukas P, Thorsteinsdottir U, Groop LC, Stefansson K, Hu F, Pankow JS, Dupuis J, Meigs JB, Altshuler D, Boehnke M, McCarthy MI: Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nature genetics 2012;44:981-990