**Figure legends**

**Figure 1.** Quantile-quantile plot of -log10 p-values from different methods for testing the causal effect either in the absence or in the presence of horizontal pleiotropic effect under null simulations. Compared methods include CoMM (green), PMR-Egger (red), TWAS (blue), LDA MR-Egger (black), SMR (orange), and PrediXcan (purple). Null simulations are performed under different horizontal pleiotropic effect sizes: (**A**) ; (**B**) ; (**C**) ; (**D**) . Only p-values from PMR-Egger adhere to the expected diagonal line across a range of horizontal pleiotropic effect sizes.

**Figure 2.** Power of different methods under various simulation scenarios. Power (y-axis) at a false discovery rate of 0.1 to detect the causal effect (**A**-**D**) or the horizontal pleiotropic effect (**E**-**F**) is plotted against different causal effect size characterized by (x-axis). Compared methods include CoMM (green), PMR-Egger (red), TWAS (blue), LDA MR-Egger (black), SMR (orange), and PrediXcan (purple). Simulations are performed under different horizontal pleiotropic effect sizes: (**A**) ; (**B**) ; (**C**, **E**) ; (**D**, **F**) .

**Figure 3.** Quantile-quantile plot of -log10 p-values from different methods for testing the horizontal pleiotropic effect either in the absence or in the presence of causal effect under null simulations. Compared methods include PMR-Egger (red), LDA MR-Egger (black), and MR-PRESSO (dodger blue). Null simulations are performed under different causal effect sizes characterized by : (**A**) ; (**B**) ; (**C**) ; and (**D**) . Only p-values from PMR-Egger adhere to the expected diagonal line across a range of horizontal pleiotropic effect sizes. Due to heavy computational burden, we are only able to run 10,000 permutations for MR-PRESSO. Therefore, the minimal p-value from MR-PRESSO is .

**Figure 4.** TWAS analysis results by different methods for traits in the WTCCC data. Compared methods include CoMM (green), PMR-Egger (red), TWAS (blue), LDA MR-Egger (black), SMR (orange), and PrediXcan (purple). (**A**) Quantile-quantile plot of -log10 p-values from different methods for testing the causal effect for an exemplary trait BD. (**B**) Quantile-quantile plot of -log10 p-values from different methods for testing the causal effect for another exemplary trait T1D. (**C**) Genomic inflation factor for testing the causal effect for each of the 7 traits by different methods. (**D**) Number of causal genes identified for each of the 7 traits by different methods. (**E**) Quantile-quantile plot of -log10 p-values from different methods for testing the horizontal pleiotropic effect for an exemplary trait BD. (**F**) Quantile-quantile plot of -log10 p-values from different methods for testing the horizontal pleiotropic effect for another exemplary trait T1D. (**G**) Genomic inflation factor for testing the horizontal pleiotropic effect for each of the 7 traits by different methods. (**H**) Number of genes identified to have significant horizontal pleiotropic effect for each of the 7 traits by different methods. For **C**, **D**, **G**, **H**, the number on the x-axis represents seven traits in order: T1D, CD, RA, BD, T2D, CAD, HT.

**Figure 5.** TWAS analysis results by different methods for traits in the GERA data. Compared methods include CoMM (green), PMR-Egger (red), TWAS (blue), LDA MR-Egger (black), SMR (orange), and PrediXcan (purple). (**A**) Quantile-quantile plot of -log10 p-values from different methods for testing the causal effect for an exemplary trait Irritable Bowel Syndrome. (**B**) Quantile-quantile plot of -log10 p-values from different methods for testing the causal effect for another exemplary trait Asthma. (**C**) Genomic inflation factor for testing the causal effect for each of the 22 traits by different methods. (**D**) Number of causal genes identified for each of the 22 traits by different methods. (**E**) Quantile-quantile plot of -log10 p-values from different methods for testing the horizontal pleiotropic effect for an exemplary trait Irritable Bowel Syndrome. (**F**) Quantile-quantile plot of -log10 p-values from different methods for testing the horizontal pleiotropic effect for another exemplary trait Asthma. (**G**) Genomic inflation factor for testing the horizontal pleiotropic effect for each of the 22 traits by different methods. (**H**) Number of genes identified to have significant horizontal pleiotropic effect for each of the 22 traits by different methods. For **C**, **D**, **G**, **H**, the number on the x-axis represents 22 traits in order: Asthma, Allergic Rhinitis, CARD, Cancers, Depressive Disorder, Dermatophytosis, T2D, Dyslipidemia, HT, Hemorrhoids, Abdominal Hernia, Insomnia, Iron Deficiency, Irritable Bowel Syndrome, Macular Degeneration, Osteoarthritis, Osteoporosis, PVD, Peptic Ulcer, Psychiatric disorders, Stress Disorders, Varicose Veins.

**Figure 6.** TWAS analysis results by different methods for traits in the UK Biobank data. Compared methods include CoMM (green), PMR-Egger (red), TWAS (blue), LDA MR-Egger (black), SMR (orange), and PrediXcan (purple). (**A**) Quantile-quantile plot of -log10 p-values from different methods for testing the causal effect for an exemplary trait BMI. (**B**) Quantile-quantile plot of -log10 p-values from different methods for testing the causal effect for another exemplary trait Platelet Count. (**C**) Genomic inflation factor for testing the causal effect for each of the 10 traits by different methods. (**D**) Number of causal genes identified for each of the 10 traits by different methods. (**E**) Quantile-quantile plot of -log10 p-values from different methods for testing the horizontal pleiotropic effect for an exemplary trait BMI. (**F**) Quantile-quantile plot of -log10 p-values from different methods for testing the horizontal pleiotropic effect for another exemplary trait Platelet Count. (**G**) Genomic inflation factor for testing the horizontal pleiotropic effect for each of the 10 traits by different methods. (**H**) Number of genes identified to have significant horizontal pleiotropic effect for each of the 10 traits by different methods. For **C**, **D**, **G**, **H**, the number on the x-axis represents 10 traits in order: Height, Platelet count, Bone mineral density, Red blood cell count, FEV1-FVC ratio, BMI, RDW, Eosinophils count, Forced vital capacity, White blood cell count.

**Figure S1.** An illustrative diagram for Mendelian randomization analysis. Mendelian randomization analysis in the TWAS setting attempts to estimate the causal effect of gene expression (x) on the trait of interest (y) in the presence of confounding factors (U) by using cis-SNPs (Z) as instrumental variables. An important requirement of Mendelian randomization analysis is to model and control for horizontal pleiotropic effects.

**Figure S2.** Quantile-quantile plot of -log10 p-values from LDA MR-Egger under null simulations. LDA MR-Egger tests for either causal effect (**A**) or horizontal pleiotropic effect (**B**). We follow the same simulation design in the original LDA MR-Egger paper to perform simulations. The simulation design assumes that the SNP covariance matrix is an AR(1) covariance structure, where we set the autocorrelation parameter to be either 0.9 (green), 0.95 (red), 0.96 (blue), 0.97 (black), 0.98 (orange), and 0.99 (purple). The inflation of LDA MR-Egger becomes apparent when the autocorrelation becomes greater than 0.9, and such inflation increases with increasing autocorrelation values. The LD decay pattern under these covariance structures are plotted together with the realistic LD pattern of the *BACE1* gene (pink). The LD pattern of the *BACE1* gene is estimated either in the eQTL data (**C**) or in the GWAS data (**D**).

**Figure S3.** Quantile-quantile plot of -log10 p-values from different methods for testing the causal effect under the null simulations, in various sparse settings where only a small proportion of SNPs are associated with the gene expression level. Compared methods include CoMM (green), PMR-Egger (red), TWAS (blue), LDA MR-Egger (black), SMR (orange), and PrediXcan (purple). Simulations are performed either in the absence (; **A**, **C**, **E**) or in the presence of horizontal pleiotropic effect (; **B**, **D**, **F**). Either one SNP (**A**, **B**), 1% of SNPs (**C**, **D**), or 10% SNPs (**E**, **F**) have non-zero effects on gene expression. Only p-values from PMR-Egger adhere to the expected diagonal line across a range of settings.

**Figure S4.** Quantile-quantile plot of -log10 p-values from different methods for testing the causal effect under the null, across different gene expression heritability values. Compared methods include CoMM (green), PMR-Egger (red), TWAS (blue), LDA MR-Egger (black), SMR (orange), and PrediXcan (purple). Simulations are performed or in the presence of horizontal pleiotropic effect () with gene expression heritability being either (**A**) or (**B**) . Only p-values from PMR-Egger adhere to the expected diagonal line across a range of settings.

**Figure S5.** Quantile-quantile plot of -log10 p-values from PMR-Egger for testing the causal effect under the null, under various sparse horizontal pleiotropic effect settings. Simulations are performed under different horizontal pleiotropic effect sizes: (**A**) ; (**B**) ; (**C**) ; (**D**) . In each panel, only a fixed proportion of SNPs (10%, 30%, or 50%) have non-zero horizontal pleiotropic effects. p-values from PMR-Egger behave well across a range of sparse horizontal pleiotropic effect settings.

**Figure S6.** Quantile-quantile plot of -log10 p-values from PMR-Egger for testing the causal effect under the null, under various directional horizontal pleiotropic effect assumptions. Simulations are performed under different horizontal pleiotropic effect sizes: (**A**) ; (**B**) ; (**C**) ; (**D**) . In each panel, a fixed proportion of SNPs (10%, 30%, or 50%) have positive horizontal pleiotropic effects while the remaining proportion of SNPs have negative horizontal pleiotropic effects. p-values from PMR-Egger behave reasonably well across a range of directional or balanced horizontal pleiotropic effect settings, except in the extreme case where horizontal pleiotropic effect size is very large () and where the effect size signs across SNPs are approximately balanced.

**Figure S7.** Power for testing causal effect by different methods in various sparse settings where only a small proportion of SNPs are associated with the gene expression level. Power (y-axis) at a false discovery rate of 0.1 to detect the causal effect is plotted against different causal effect size characterized by (x-axis). Compared methods include CoMM (green), PMR-Egger (red), TWAS (blue), LDA MR-Egger (black), SMR (orange), and PrediXcan (purple). Simulations are performed either in the absence (; **A**, **C**, **E**) or in the presence of horizontal pleiotropic effect (; **B**, **D**, **F**). Either one SNP (**A**, **B**), 1% of SNPs (**C**, **D**), or 10% SNPs (**E**, **F**) have non-zero effects on gene expression.

**Figure S8.** Boxplot displays causal effect estimates by PMR-Egger in the absence or presence of horizontal pleiotropic effect. Simulations are performed under different horizontal pleiotropic effect sizes (x-axis: or ). For each horizontal pleiotropic effect size, we examined three true causal effect sizes (red), 0.2 (green), or 0.245 (purple), which corresponds to , 0.4% and 0.6%, respectively. The horizontal red dashed lines represent the three true values of . PMR-Egger produces approximately unbiased causal effect size estimates across different scenarios. 10000 replicates are included for each simulation scenario.

**Figure S9.** Quantile-quantile plot of -log10 p-values from different methods for testing the horizontal pleiotropic effect under null simulations, in various sparse settings where only a small proportion of SNPs are associated with the gene expression level. Compared methods include PMR-Egger (red) and LDA MR-Egger (black). Simulations are performed either in the absence (; **A**, **C**, **E**) or in the presence of causal effect (; **B**, **D**, **F**). Either one SNP (**A**, **B**), 1% of SNPs (**C**, **D**), or 10% SNPs (**E**, **F**) have non-zero effects on gene expression. Only p-values from PMR-Egger adhere to the expected diagonal line across a range of settings. Note that we do not include MR-PRESSO into comparison here due to the relatively higher computation burden and it is difficult in MR-PRESSO to pre-specify the number of simulated expected distribution and obtain the exact p values.

**Figure S10.** Power of different methods for identifying horizontal pleiotropic effect, in various sparse settings where only a small proportion of SNPs are associated with the gene expression level. Compared methods include PMR-Egger (red) and LDA MR-Egger (black). Power (y-axis) at a false discovery rate of 0.1 to detect the horizontal pleiotropic effect is plotted against sparsity levels, either in the absence (; **A**) or in the presence of causal effect (; **B**). In terms of sparsity level (x-axis), either one SNP, 1% SNPs, 10% SNPs and 100% SNPs have non-zero effects on gene expression.

**Figure S11.** Power of different methods for identifying horizontal pleiotropic effect under various model misspecifications of the horizontal pleiotropic effect. Compared methods include PMR-Egger (red) and LDA MR-Egger (black). On both panels, power (y-axis) to detect the horizontal pleiotropic effect is measured at a false discovery rate of 0.1 for a fixed horizontal pleiotropic effect size . (**A**) Power is plotted against the proportion of SNPs displaying non-zero horizontal pleiotropy effects (10%, 30%, or 50%). (**B**) Power is plotted against the proportion of SNPs exhibiting negative horizontal pleiotropy effects (10%, 30%, or 50%), whereas the remaining proportion of SNPs exhibiting positive effects.

**Figure S12.** Boxplot displays horizontal pleiotropic effect estimates by PMR-Egger in the absence or presence of causal effect. Simulations are performed under different causal effect sizes (x-axis: or ). For each causal effect size, we examined three true horizontal pleiotropic effect sizes (red), 0.0005 (green), or 0.001 (purple). The horizontal red dashed lines represent the three true values of . PMR-Egger produces approximately unbiased causal effect size estimates across different scenarios. 10,000 replicates are included for each simulation scenario.

**Figure S13.** Quantile-quantile plot of -log10 p-values from different methods in the TWAS application to WTCCC. Compared methods include CoMM (green), PMR-Egger (red), TWAS (blue), LDA MR-Egger (black), SMR (orange), and PrediXcan (purple). Top panels: Quantile-quantile plot of -log10 p-values from different methods for testing the causal effect for five traits (CAD, CD, HT, RA and T2D). Bottom panels: Quantile-quantile plot of -log10 p-values from different methods for testing the horizontal pleiotropic effect for the same five traits.

**Figure S14.** Quantile-quantile plot of -log10 p-values from different methods for testing causal effect in the TWAS application to 20 traits in GERA. Compared methods include CoMM (green), PMR-Egger (red), TWAS (blue), LDA MR-Egger (black), SMR (orange), and PrediXcan (purple).

**Figure S15.** Quantile-quantile plot of -log10 p-values from different methods for testing causal effect in the TWAS application to 8 traits in UK Biobank. Compared methods include CoMM (green), PMR-Egger (red), TWAS (blue), LDA MR-Egger (black), SMR (orange), and PrediXcan (purple).

**Figure S16.** Overlap of genes detected by PMR-Egger and CoMM. (**A**) Jaccard index measures relatively high overlap between genes detected by PMR-Egger and genes detected by CoMM. (**B**) Mean of the estimated for the set of genes that are detected by both CoMM and PMR-Egger (“overlapped”), and for the set of genes that only detected by CoMM (“CoMM only”), across 13 traits that have at least 2 detected associations. In (B), p-values are calculated based on Wilcoxon rank sum test and the traits on x-axis are: Ast, Asthma; Dys, Dyslipidemia; MD, Macular Degeneration; PC, Platelet count; BMD, Bone mineral density; RBC, Red blood cell count; FFC, FEV1-FVC ratio; EC, Eosinophils count; FVC, Forced vital capacity; WBC, White blood cell count.

**Figure S17.** Quantile-quantile plot of -log10 p-values from different methods for testing horizontal pleiotropic effect in the TWAS application to 20 traits in GERA. Compared methods include CoMM (green), PMR-Egger (red), TWAS (blue), LDA MR-Egger (black), SMR (orange), and PrediXcan (purple).

**Figure S18.** Quantile-quantile plot of -log10 p-values from different methods for testing horizontal pleiotropic effect in the TWAS application to 8 traits in UK Biobank. Compared methods include CoMM (green), PMR-Egger (red), TWAS (blue), LDA MR-Egger (black), SMR (orange), and PrediXcan (purple).

**Figure S19.** Scatter plot of -log10 p-values by PMR-Egger for testing causal effect versus -log10 p-values by PMR-Egger for testing horizontal pleiotropic effect across genes for each of four traits in WTCCC. Each dot represents one gene. Only traits with at least one gene that has either a significant causal effect or a significant horizontal pleiotropic effect are displayed.

**Figure S20.** Scatter plot of -log10 p-values by PMR-Egger for testing causal effect versus -log10 p-values by PMR-Egger for testing horizontal pleiotropic effect across genes for each of 10 traits in GERA. Each dot represents one gene. Only traits with at least one gene that has either a significant causal effect or a significant horizontal pleiotropic effect are displayed.

**Figure S21.** Scatter plot of -log10 p-values by PMR-Egger for testing causal effect versus -log10 p-values by PMR-Egger for testing horizontal pleiotropic effect across genes for each of 10 traits in UK Biobank. Each dot represents one gene. Only traits with at least one gene that has either a significant causal effect or a significant horizontal pleiotropic effect are displayed.

**Figure S22.** Quantile-quantile plot of -log10 p-values for testing the causal effect in null simulations, under a variance component modeling assumption for the horizontal pleiotropic effects. In simulations, each follows the normal distribution with variance equals 0.001/556. The p-values are calibrated if we knew the true hyper-parameters (**A**). However, due to the uncertainty in hyper-parameter estimates, p-values become are overly conservative when we estimate hyper-parameters (**B**).

**Table legends**

**Table S1.** Genomic inflation factor for different methods in analyzing the three GWAS data sets.

**Table S2.** Number of significant genes identified by different methods in the three GWAS data sets under Bonferroni correction.