Supporting information for:

**Biomarker discovery for MASLD utilizing Mendelian Randomization, Machine Learning, and external validation**

**Authors**:

Gong Feng1, Giovanni Targher2,3, Christopher D. Byrne4, Na He5, Man Mi6, Yi Liu7, Hongbin Zhu8, Ming-Hua Zheng9,10, Feng Ye1\*

**SUPPLEMENTARY MATERIAL**

[Supplementary Method 1 5](#_Toc201419368)

[Supplementary Method 2 6](#_Toc201419369)

[Supplementary Method 3 7](#_Toc201419370)

[Supplementary Method 4 8](#_Toc201419374)

[Supplementary Method 5 9](#_Toc201419375)

[Table S1. Causal relationships between the six proteins (exposure) and MASLD (outcome). 10](#_Toc201419376)

[Table S2. Heterogeneity and pleiotropy analysis between the six proteins (explore) and MASLD (outcome). 12](#_Toc201419377)

[Table S3. Baseline characteristics of the validation cohort used for serum total protein–related analysis. 13](#_Toc201419378)

[Table S4. Adjusted odds ratios of the serum total protein levels for the risk of MASLD in the validation cohort. 15](#_Toc201419379)

[Table S5. Adjusted hazard ratios of the serum total protein levels for the risk of all-cause and cause-specific mortality among adult individuals in the United States from the NHANES 1999-2006 database. 16](#_Toc201419380)

[Figure S1. Causal relationship between APOE (exposure) and MASLD (outcome). 17](#_Toc201419381)

[Figure S2. Causal relationship between CNPY4 (exposure) and MASLD (outcome). 18](#_Toc201419382)

[Figure S3. Causal relationship between ENTPD6 (exposure) and MASLD (outcome). 19](#_Toc201419383)

[Figure S4. Causal relationship between HLA-A (exposure) and MASLD (outcome). 20](#_Toc201419384)

[Figure S5. Causal relationship between SCG3 (exposure) and MASLD (outcome). 21](#_Toc201419385)

[Figure S6. Scatterplots of the MR analysis between Apolipoprotein A (exposure), GGT (exposure), and MASLD (outcome). 22](#_Toc201419386)

[Figure S7. Scatterplots of the MR analysis between albumin (exposure), HDL-C (exposure), and MASLD (outcome). 23](#_Toc201419387)

[Figure S8. Scatterplots of the MR analysis between IGF-1 (exposure), urinary sodium (exposure) and MASLD (outcome). 24](#_Toc201419388)

[Figure S9. Scatterplots of the MR analysis between triglycerides (exposure) and MASLD (outcome). 25](#_Toc201419389)

[Figure S10. Causal relationship between HLA-A (exposure) and serum total protein levels (outcome). 26](#_Toc201419390)

[Figure S11. Bootstrap ROC curves for the random forest model in the training and validation sets 27](#_Toc201419391)

[Figure S12. Differential expression analysis of CNPY4 in hepatocellular carcinoma and other cancer types. 28](#_Toc201419392)

[Figure S13. Association between CNPY4 expression and overall survival in hepatocellular carcinoma and other cancer types. 29](#_Toc201419393)

[Figure S14. Correlation between CNPY4 expression and immune cell infiltration in hepatocellular carcinoma and other cancers. 30](#_Toc201419394)

[Figure S15. Differential expression analysis of ENTPD6 in hepatocellular carcinoma and other cancer types. 31](#_Toc201419395)

[Figure S16. Association between ENTPD6 expression and overall survival in hepatocellular carcinoma and other cancer types. 32](#_Toc201419396)

[Figure S17. Correlation between ENTPD6 expression and immune cell infiltration in hepatocellular carcinoma and other cancers. 33](#_Toc201419397)

# Supplementary Method 1

Specially, when screening the instrumental variables for 35 clinical biomarkers, we used a genome-wide significance threshold (*P* < 5×10⁻⁸). Additionally, we set the single-nucleotide polymorphism (SNP)-specific threshold of the coefficient of determination r² below 0.001, and the width of the linkage disequilibrium region to be greater than 10,000 kilobases (kb). We calculated the F value of the instrumental variable SNPs and excluded the instrumental variables with an F statistic less than 10 to mitigate the bias caused by weak instrumental variables.

# Supplementary Method 2

To ensure the reliability of the MR analysis results, we undertook heterogeneity, pleiotropy, and sensitivity analyses. The Cochrane Q statistic was used to evaluate the heterogeneity of the instrumental variables, with P<0.05 indicating the presence of heterogeneity. The pleiotropy test was performed through the intercept term of the MR-Egger regression and MR-PRESSO, where MR-PRESSO was utilized to detect and correct outliers caused by horizontal pleiotropy. Furthermore, the sensitivity analysis was conducted using the 'leave-one-out' method to exclude the influence of individual SNPs on the results, ensuring that no single SNP would significantly affect the causal relationships between proteins, biomarkers, and MASLD.

# Supplementary Method 3

We collected serum samples from 415 participants at the First Affiliated Hospital of Xi'an Medical University, including 330 MAFLD patients and 85 healthy control subjects. The diagnosis of MAFLD was performed using the FibroScan 502 device (Echosens, Paris, France) within one month of the clinic visit and blood tests. Experienced operators, trained by the manufacturer or certified delegates, carried out the controlled attenuation parameter (CAP) and liver stiffness measurements. Patients with a CAP value ≥ 248 dB/m, as assessed by FibroScan, were considered to have hepatic steatosis. Fibroscan has already been recommended by major clinical practice guidelines and adopted by numerous studies related to MAFLD.1 This part study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the First Affiliated Hospital of Xi'an Medical University (approval number: XYFY2018LSK-003).

Reference

[1] Lin H, Lee HW, Yip TC, Tsochatzis E, Petta S, Bugianesi E, Yoneda M, Zheng MH, Hagström H, Boursier J, Calleja JL, Goh GB, Chan WK, Gallego-Durán R, Sanyal AJ, de Lédinghen V, Newsome PN, Fan JG, Castéra L, Lai M, Harrison SA, Fournier-Poizat C, Wong GL, Pennisi G, Armandi A, Nakajima A, Liu WY, Shang Y, de Saint-Loup M, Llop E, Teh KK, Lara-Romero C, Asgharpour A, Mahgoub S, Chan MS, Canivet CM, Romero-Gomez M, Kim SU, Wong VW; VCTE-Prognosis Study Group. Vibration-Controlled Transient Elastography Scores to Predict Liver-Related Events in Steatotic Liver Disease. JAMA. 2024 Apr 16;331(15):1287-1297. doi: 10.1001/jama.2024.1447. PMID: 38512249; PMCID: PMC10958386.

# Supplementary Method 4

By integrating data from the Genotype-Tissue Expression (GTEx) database and The Cancer Genome Atlas (TCGA), we systematically analyzed the expression differences of these molecular biomarkers between primary tumor tissues (particularly LIHC) and normal tissues. Subsequently, we employed the Cox proportional hazards regression model to explore the associations between the expression levels of these molecular biomarkers and overall survival (OS) prognosis across various cancer types, with a specific emphasis on LIHC. Furthermore, to gain deeper insights into the potential mechanisms of action of key proteins in the tumor microenvironment, we examined the correlations between the expression levels of molecular biomarkers and the infiltration degrees of different immune cells by integrating immune infiltration-related matrix data. The outcomes of the correlation analysis were visualized using heatmaps, thereby uncovering the potential pathways through which they may regulate the tumor immune microenvironment.

# Supplementary Method 5

The initial study population consisted of 41,474 participants from the USA in the four-cycle NHANES (1999-2006) database. Detailed information about the survey design and methods is available on the NHANES website [Centers for Disease Control and Prevention (CDC), http://cdc.gov/nchs/nhanes)]. Mortality outcomes were determined through linkage with the National Death Index (NDI), covering deaths that occurred between January 1, 1999, and December 31, 20152. Cardiovascular mortality was defined as death resulting from heart disease or cerebrovascular disease. Time-to-death was calculated from baseline to the date of death or the end of follow-up on December 31, 2015, whichever occurred first.

**References**

1. Niezen S, Mehta M, Jiang ZG, et al. Coffee Consumption Is Associated With Lower Liver Stiffness: A Nationally Representative Study. Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association. 2022;20(9):2032-40.e6. Epub 2021/10/10.

2. Nguyen VK, Colacino J, Chung MK, et al. Characterising the relationships between physiological indicators and all-cause mortality (NHANES): a population-based cohort study. Lancet Healthy Longev. 2021;2(10):e651-e62. Epub 20210920.

3. Konyn P, Alshuwaykh O, Dennis BB, et al. Gallstone Disease and Its Association With Nonalcoholic Fatty Liver Disease, All-Cause and Cause-Specific Mortality. Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association. 2023;21(4):940-8 e2. Epub 2022/06/02.

|  |  |  |  |
| --- | --- | --- | --- |
| Table S1. Causal relationships between six proteins (exposure) and MASLD (outcome). | | | |
| Proteins | Methods | Number of SNPs | OR (95%CI) | *P* |
| APOE |  |  |  |  |
|  | MR Egger | 22 | 1.079 (1.029 - 1.131) | 5.38E-03 |
|  | Weighted median | 22 | 1.041 (1.003 - 1.081) | 3.50E-02 |
|  | Inverse variance weighted | 22 | 1.057 (1.031 - 1.083) | 4.62E-03 \* |
|  | Simple mode | 22 | 1.030 (0.953 - 1.114) | 4.63E-01 |
|  | Weighted mode | 22 | 1.114 (1.041 - 1.191) | 5.06E-03 |
| CNPY4 |  |  |  |  |
|  | MR Egger | 22 | 1.118 (1.049 - 1.193) | 2.79E-03 |
|  | Weighted median | 22 | 1.073 (1.041 - 1.107) | 5.70E-06 |
|  | Inverse variance weighted | 22 | 1.054 (1.029 - 1.081) | 8.38E-03 \* |
|  | Simple mode | 22 | 1.090 (1.025 - 1.160) | 1.21E-02 |
|  | Weighted mode | 22 | 1.087 (1.026 - 1.152) | 9.69E-03 |
| ENTPD6 |  |  |  |  |
|  | MR Egger | 36 | 1.045 (1.024 - 1.066) | 1.64E-04 |
|  | Weighted median | 36 | 1.034 (1.014 - 1.054) | 8.59E-04 |
|  | Inverse variance weighted | 36 | 1.031 (1.016 - 1.045) | 5.75E-03 \* |
|  | Simple mode | 36 | 1.026 (0.990 - 1.062) | 1.70E-01 |
|  | Weighted mode | 36 | 1.040 (1.015 - 1.065) | 3.16E-03 |
| HLA-A |  |  |  |  |
|  | MR Egger | 101 | 0.957 (0.930 - 0.984) | 2.58E-03 |
|  | Weighted median | 101 | 0.975 (0.960 - 0.990) | 9.20E-04 |
|  | Inverse variance weighted | 101 | 0.969 (0.960 - 0.979) | 1.81E-06 \* |
|  | Simple mode | 101 | 0.988 (0.943 - 1.035) | 6.14E-01 |
|  | Weighted mode | 101 | 0.999 (0.961 - 1.038) | 9.42E-01 |
| SCG3 |  |  |  |  |
|  | MR Egger | 23 | 0.953 (0.912 - 0.997) | 4.91E-02 |
|  | Weighted median | 23 | 0.947 (0.922 - 0.972) | 4.63E-05 |
|  | Inverse variance weighted | 23 | 0.956 (0.937 - 0.975) | 4.62E-03 \* |
|  | Simple mode | 23 | 0.949 (0.906 - 0.994) | 3.71E-02 |
|  | Weighted mode | 23 | 0.944 (0.910 - 0.979) | 5.56E-03 |
| TOR1AIP1 |  |  |  |  |
|  | MR Egger | 43 | 0.941 (0.906 - 0.977) | 2.64E-03 |
|  | Weighted median | 43 | 0.957 (0.934 - 0.981) | 4.28E-04 |
|  | Inverse variance weighted | 43 | 0.964 (0.950 - 0.979) | 8.13E-04 \* |
|  | Simple mode | 43 | 0.948 (0.910 - 0.988) | 1.53E-02 |
|  | Weighted mode | 43 | 0.953 (0.929 - 0.978) | 6.06E-04 |

*Note*: \* indicates that the IVW was used, and the p-value has been adjusted using FDR correction.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table S2. Heterogeneity and pleiotropy analysis between six proteins (exposures) and MASLD (outcome). | | | | | | | | |
|  | MR-Egger regression | | IVW | | |  | MR-Egger | |
|  | Intercept | *P* |  | Q | Q\_ *P* value |  | Q | Q\_ *P* value |
| APOE | -0.014 | 0.332 |  | 30.224 | 0.088 |  | 28.800 | 0.092 |
| CNPY4 | -0.026 | 0.068 |  | 25.891 | 0.211 |  | 21.838 | 0.349 |
| ENTPD6 | -0.013 | 0.079 |  | 32.518 | 0.588 |  | 29.230 | 0.701 |
| HLA-A | 0.007 | 0.330 |  | 96.408 | 0.583 |  | 95.448 | 0.582 |
| SCG3 | 0.002 | 0.898 |  | 5.542 | 1.000 |  | 5.525 | 1.000 |
| TOR1AIP1 | 0.014 | 0.169 |  | 42.483 | 0.450 |  | 40.526 | 0.492 |

# Table S3. Baseline characteristics of the validation cohort used for serum total protein–related analysis.

|  |  |  |
| --- | --- | --- |
| Characteristics | Control (N=85) | MASLD (N=330) |
| **Demographics** |  |  |
| Men, n (%) | 38 (44.7%) | 220 (66.7%) |
| Age (years) | 56 (47, 62) | 43 (32, 53) |
| **Concomitant diseases** |  |  |
| Diabetes, n (%) | 12 (14.1%) | 72 (21.8%) |
| Hypertension, n (%) | 16 (18.8%) | 70 (21.2%) |
| **Body measurements** |  |  |
| BMI, kg/m2 | 22.8 (21.28, 24.28) | 26.28 (24.1, 28.0) |
| **Laboratory parameters** |  |  |
| AST, (U/L) | 21.0 (18.1, 25.4) | 34.2 (24.3, 51.9) |
| ALT, (U/L) | 15.9 (13, 24) | 46.3 (27.1, 84.8) |
| HbA1c, % | 6.03±1.82 | 6.03±1.25 |
| Total protein, g/L | 67.4 (62.7, 72.4) | 76.2 (70.9, 80.0) |
| Albumin, g/L | 42.71±4.69 | 45.72±4.72 |
| Total cholesterol, mmol/L) | 4.88±1.26 | 5.11±1.20 |
| HDL, mmol/L | 1.24±0.30 | 1.04±0.23 |
| LDL, mmol/L | 2.97±0.86 | 3.04±0.93 |
| TG, mmol/L | 1.23 (0.91, 1.94) | 1.96 (1.36, 2.91) |
| Platelet, ×109/L | 210.25±61.96 | 241.40±58.37 |
| Creatinine, µmol/L | 64.18±10.62 | 66.86±13.53 |

Abbreviations: MASLD, Metabolic dysfunction-associated steatotic liver disease; BMI, body mass index; HDL, high density lipoprotein; LDL, low-density lipoprotein; TG, triglyceride.

# Table S4. Adjusted odds ratios of the serum total protein levels for risk of MASLD in the validation cohort.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Age & sex adjusted Model | | Multivariable Model 1 | Multivariable Model 2 | | |
|  | OR (95% CI) | *P* | OR (95% CI) | *P* | OR (95% CI) | *P* |
|  |  |  |  |  |  |  |
| Total protein | 1.103 (1.064-1.143) | <0.001 | 1.092 (1.052-1.134) | < 0.001 | 1.080 (1.011-1.155) | 0.023 |

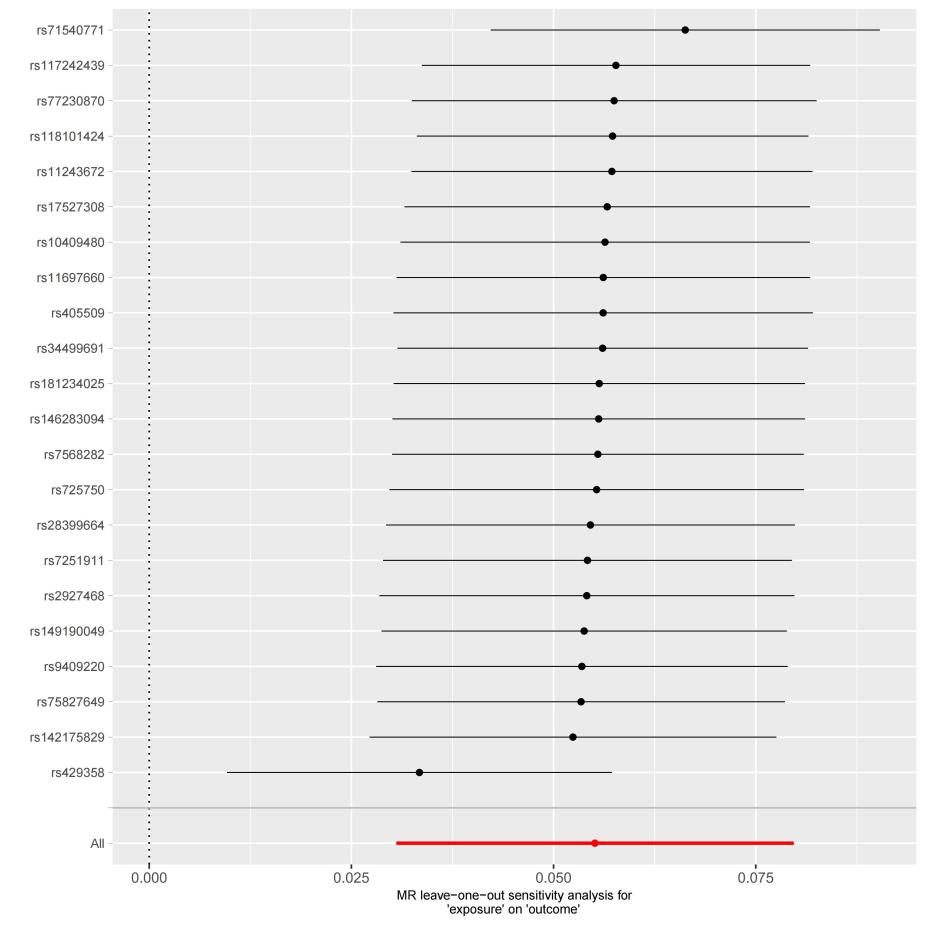
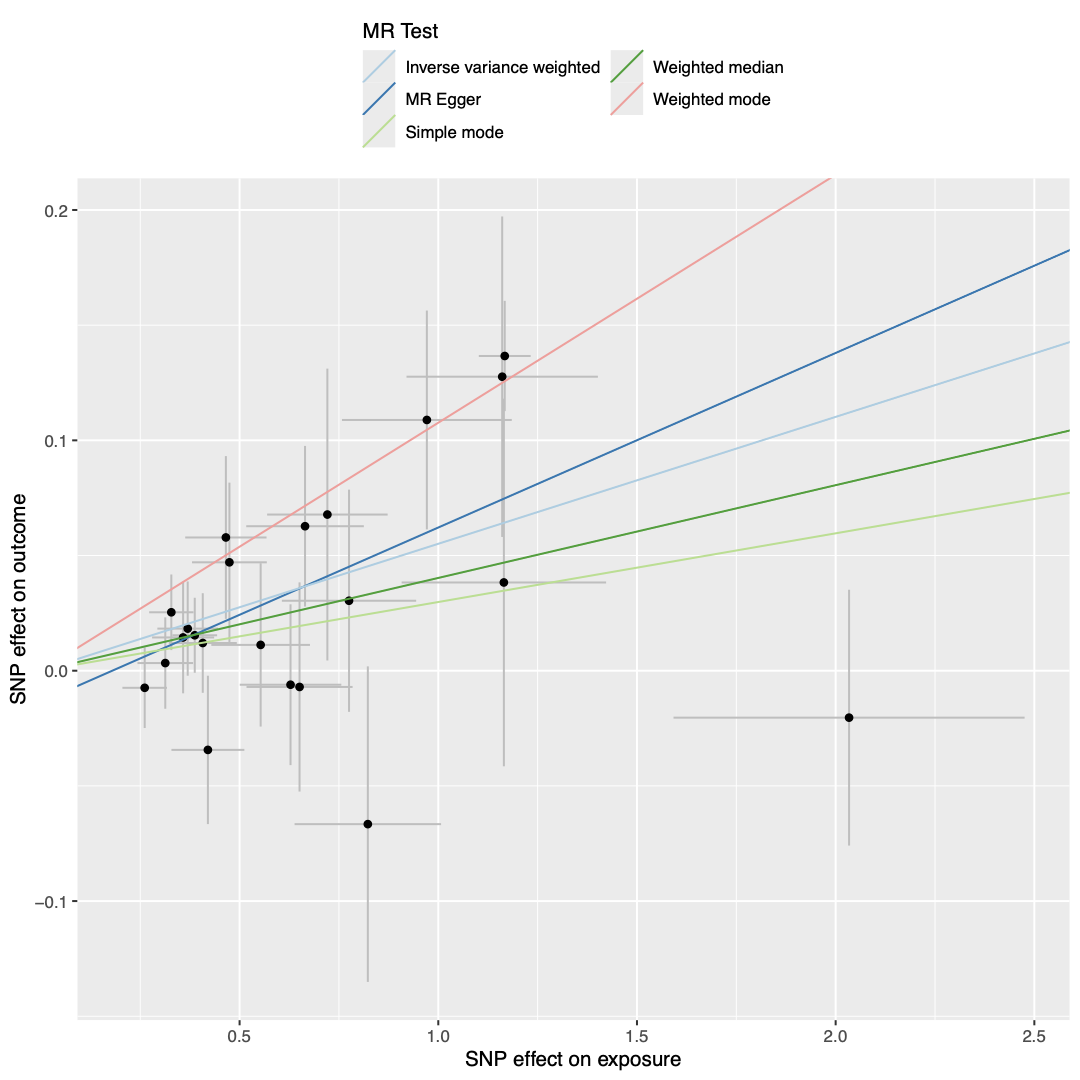
***Note****:* Multivariable regression model 1 was adjusted for age, sex, hypertension, T2DM, and body mass index. Model 2 was adjusted for model 1’s covariates *plus* serum AST, ALT, total cholesterol, HDL, LDL, triglycerides, creatinine, HbA1c, and platelet count.*Abbreviations*: ALT: alanine aminotransferase; CI, confidence interval; OR: odds ratios.

# Table S5. Adjusted hazard ratios of the serum total protein levels for risk of all-cause and cause-specific mortality among adult individuals in the United States from the NHANES 1999-2006 database.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Age-sex adjusted Model** | | **Multivariable Model 1** | | **Multivariable Model 2** | |
|  | **HR (95% CI)** | ***p*** | **HR (95% CI)** | ***p*** | **HR (95% CI)** | ***p*** |
| **All-cause mortality** |  |  |  |  |  |  |
| Total protein≥60 g/L | Ref. |  | Ref. |  | Ref. |  |
| Total protein<60 g/L | 2.180 (1.188-4.002) | 0.012 | 2.769 (1.441-5.319) | 0.002 | 2.495 (1.224-5.087) | 0.011 |
| **Cardiovascular mortality** |  |  |  |  |  |  |
| Total protein≥60 g/L | Ref. |  | Ref. |  | Ref. |  |
| Total protein<60 g/L | 3.119 (1.183-8.221) | 0.021 | 3.314 (1.051-10.451) | 0.041 | 2.926 (0.961-8.908) | 0.059 |

Note*:* Multivariable Cox regression model 1 was adjusted for age, sex, marital status, hypertension, T2DM, and body mass index. Model 2 was adjusted for model 1’s covariates *plus* serum GGT, ALT, total cholesterol, triglycerides, and platelet count.Abbreviations: ALT: alanine aminotransferase; CI, confidence interval; HR, hazard ratio; GGT: gamma-glutamyl transferase.

# Figure S1. Causal relationship between APOE (exposure) and MASLD (outcome).



**(B)**

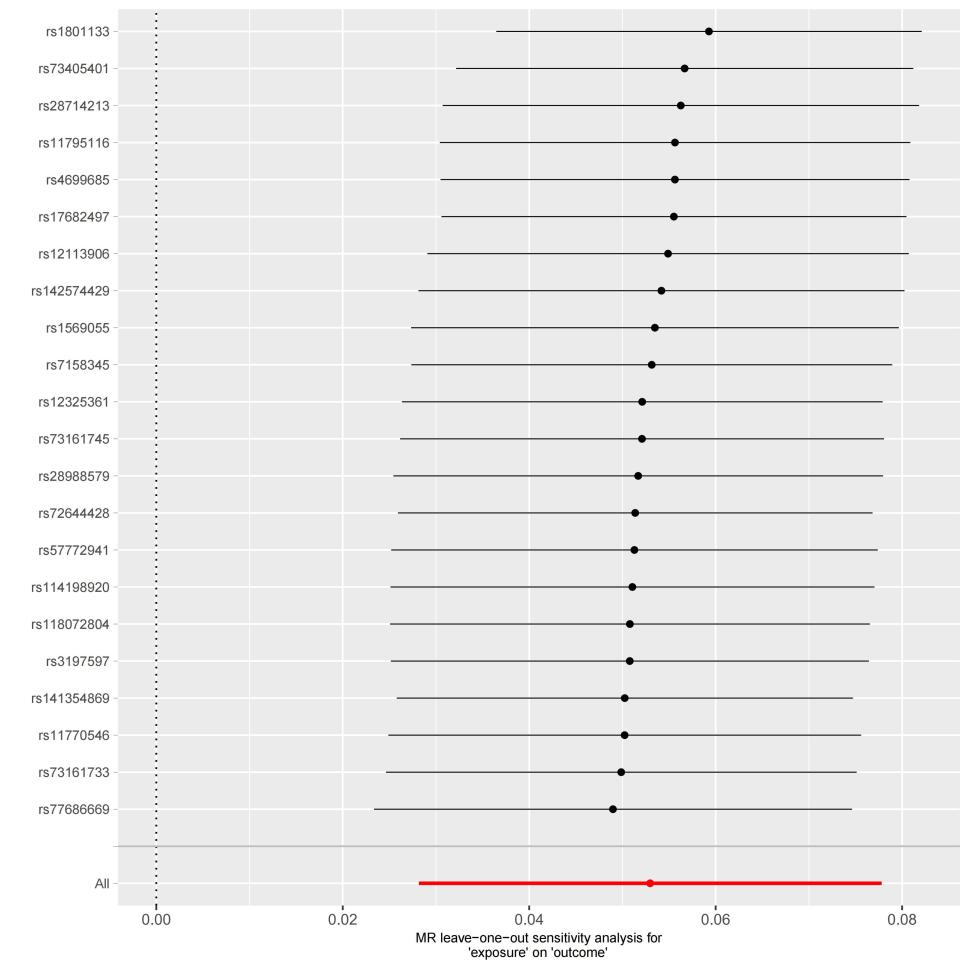
**(A)**

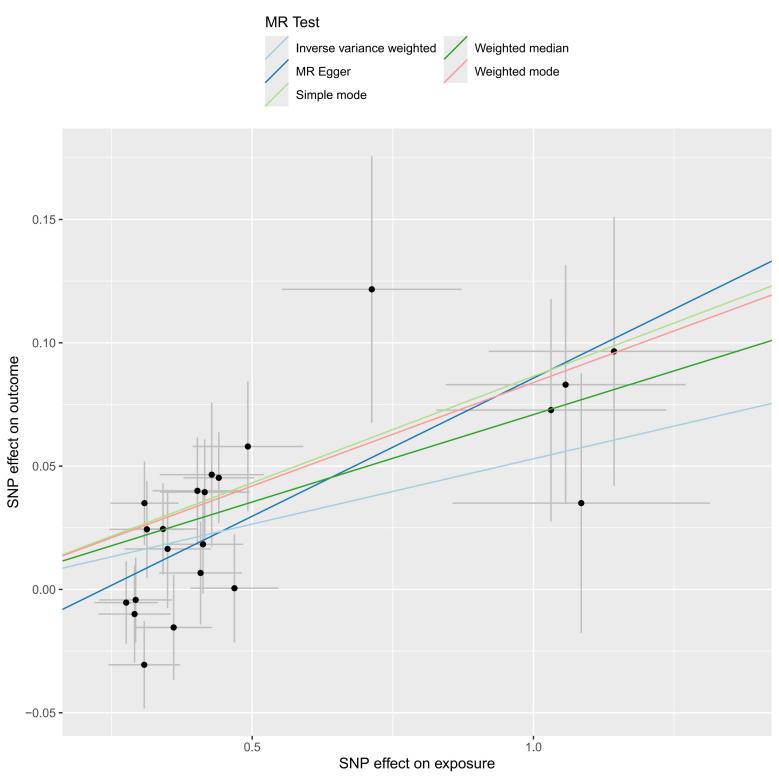
(A) scatterplot of the MR analysis; (B) sensitivity analysis

# Figure S2. Causal relationship between CNPY4 (exposure) and MASLD (outcome).

**(B)**

**(A)**

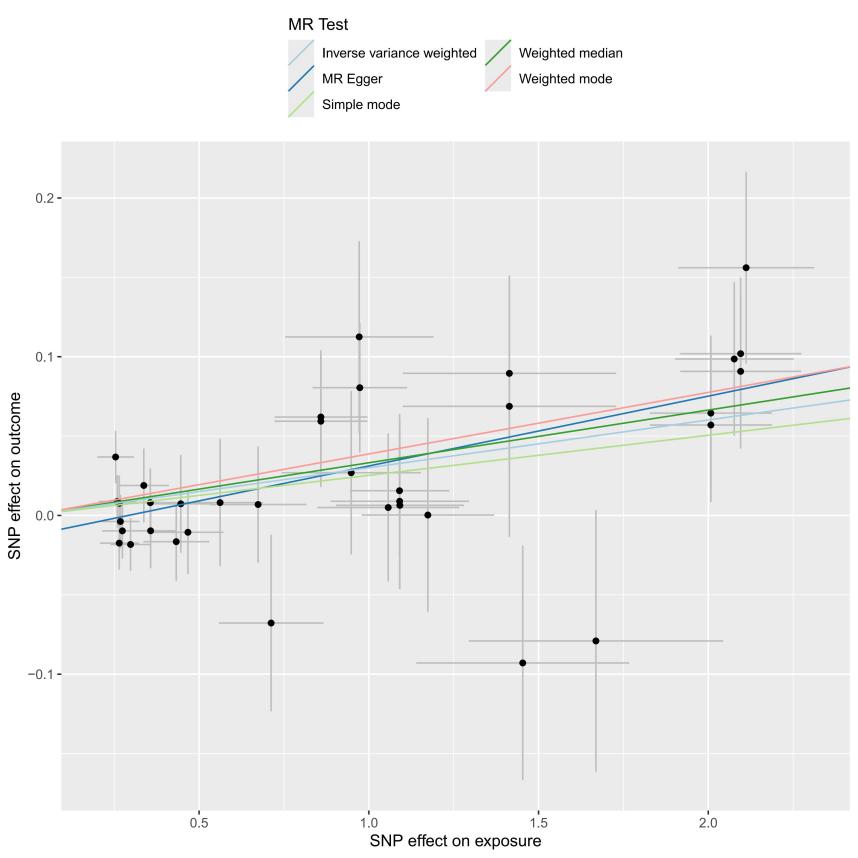




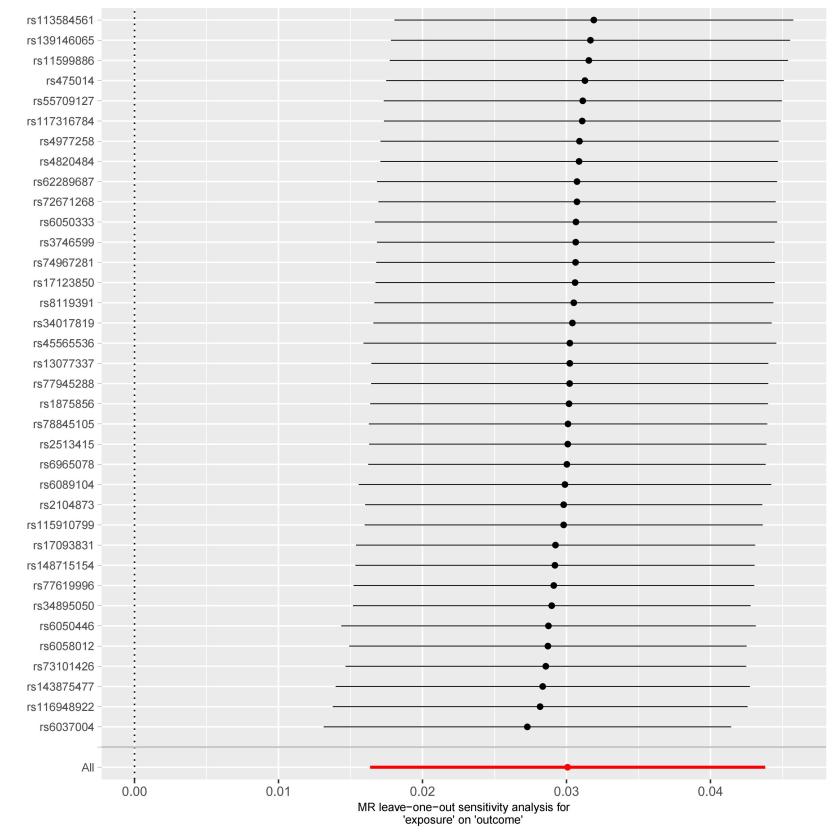
(A) scatterplot of the MR analysis; (B) sensitivity analysis

# Figure S3. Causal relationship between ENTPD6 (exposure) and MASLD (outcome).

**(A)**

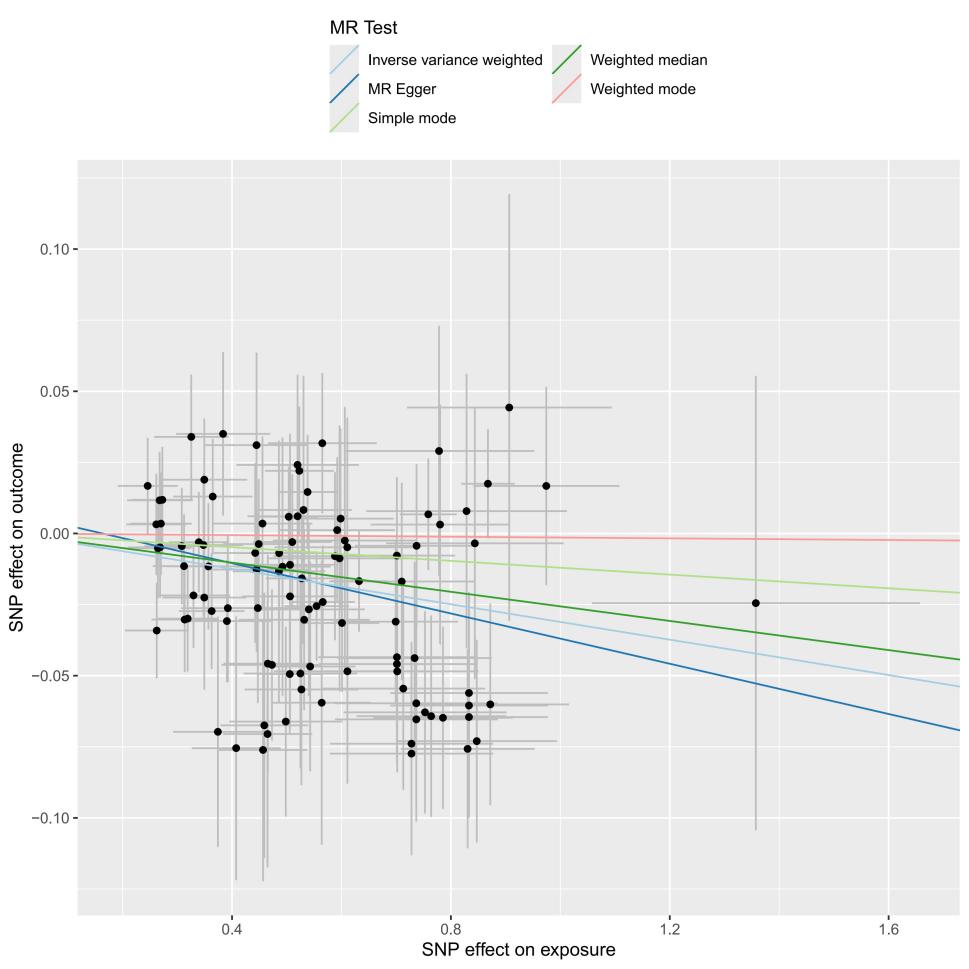
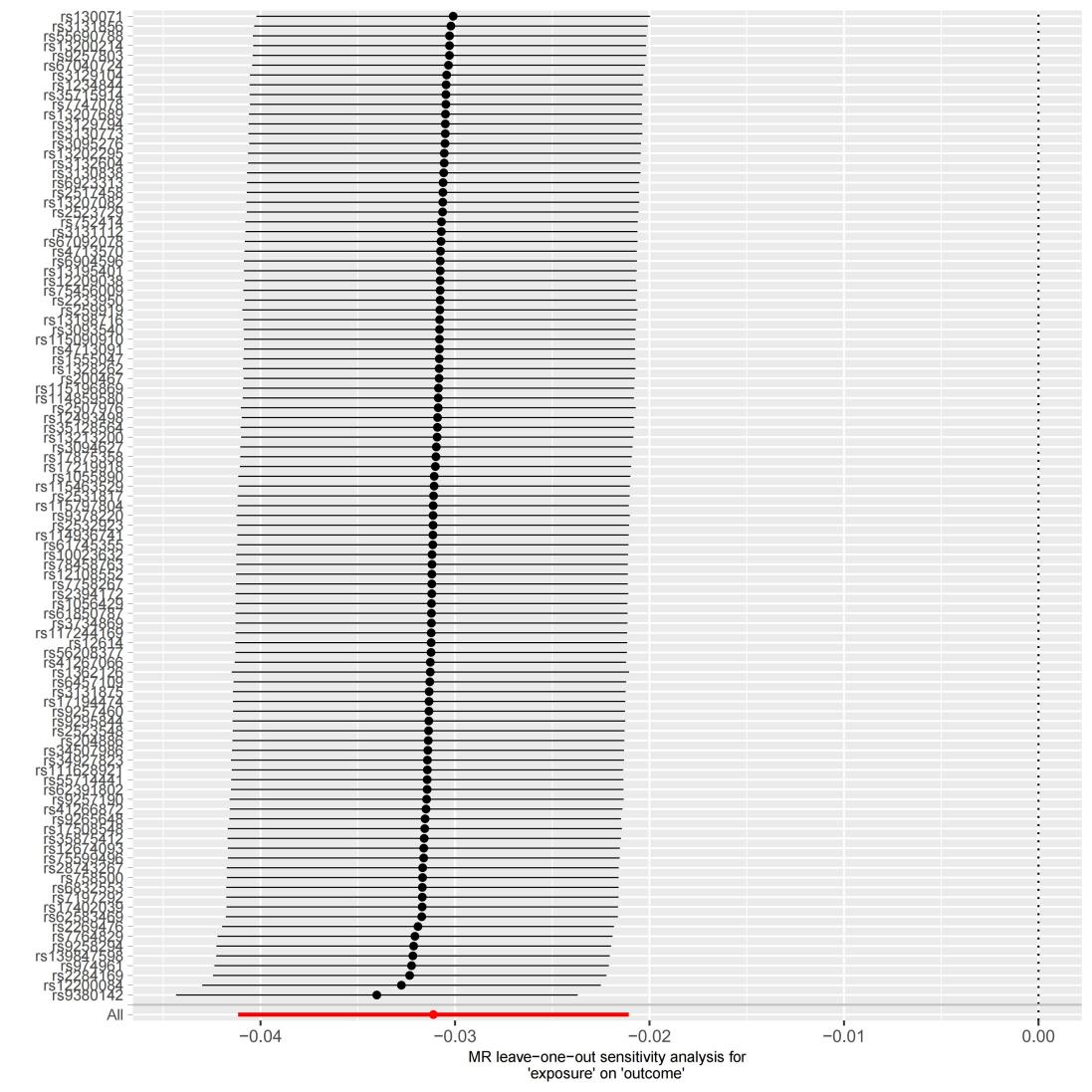


**(B)**



(A) scatterplot of the MR analysis; (B) sensitivity analysis

# Figure S4. Causal relationship between HLA-A (exposure) and MASLD (outcome).



**(A)**

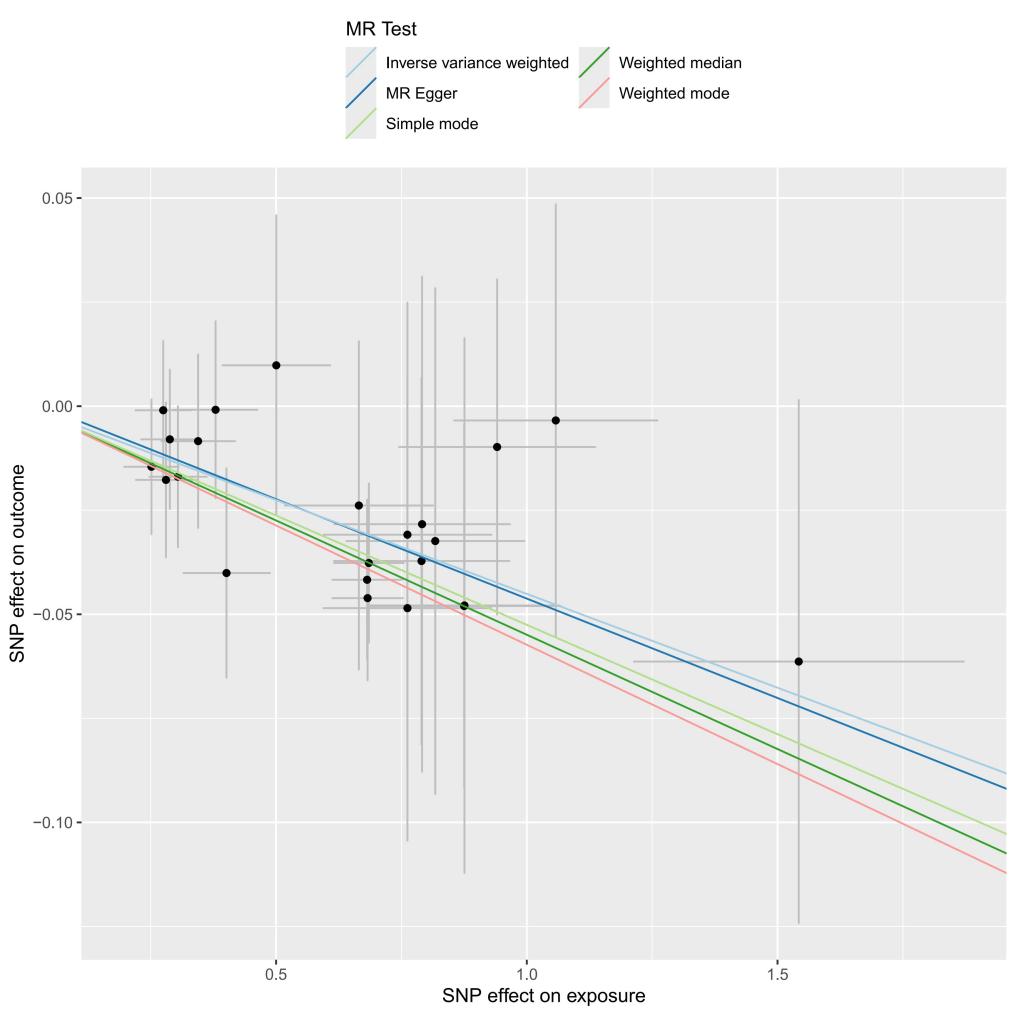
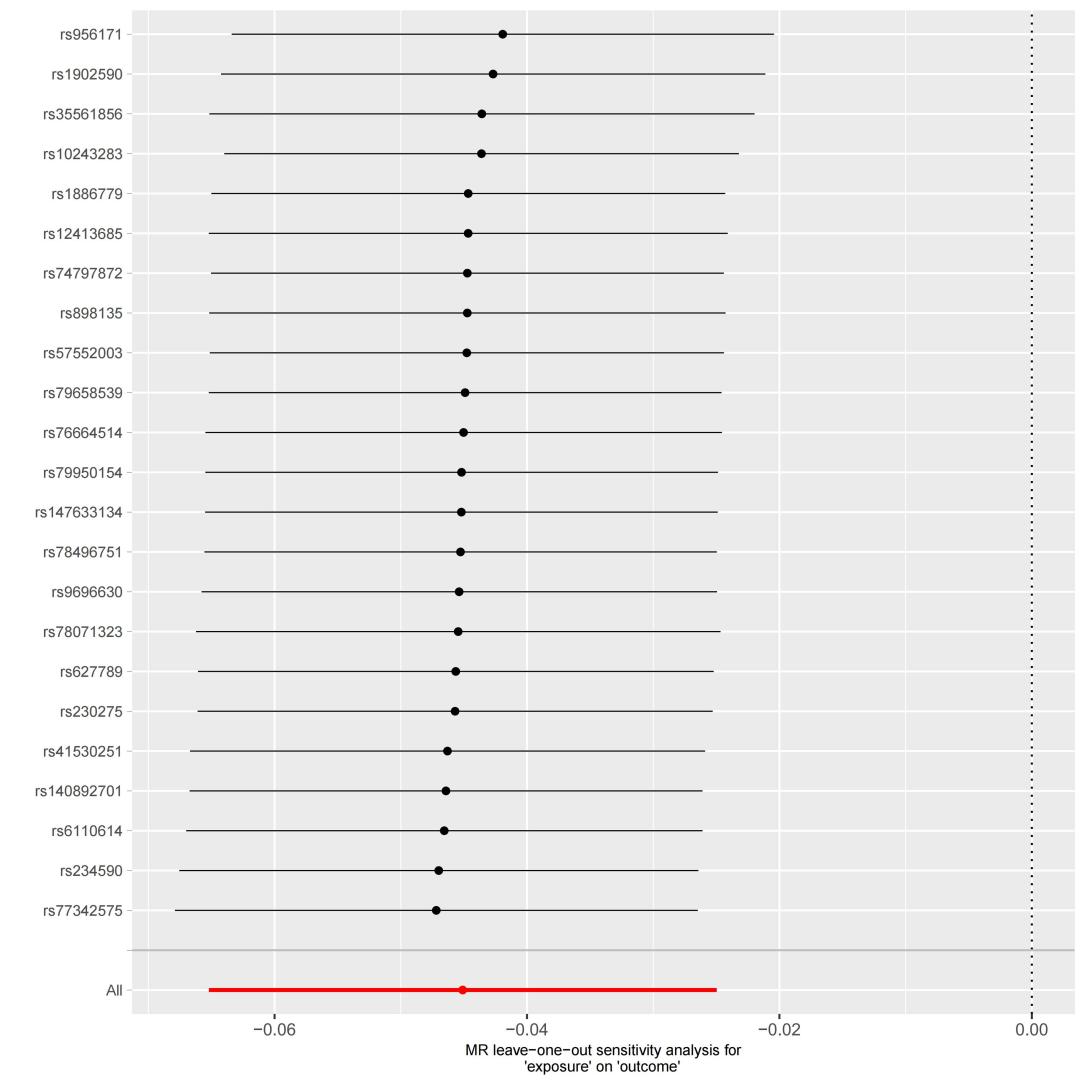
**(B)**

(A) scatterplot of the MR analysis; (B) sensitivity analysis

# Figure S5. Causal relationship between SCG3 (exposure) and MASLD (outcome).

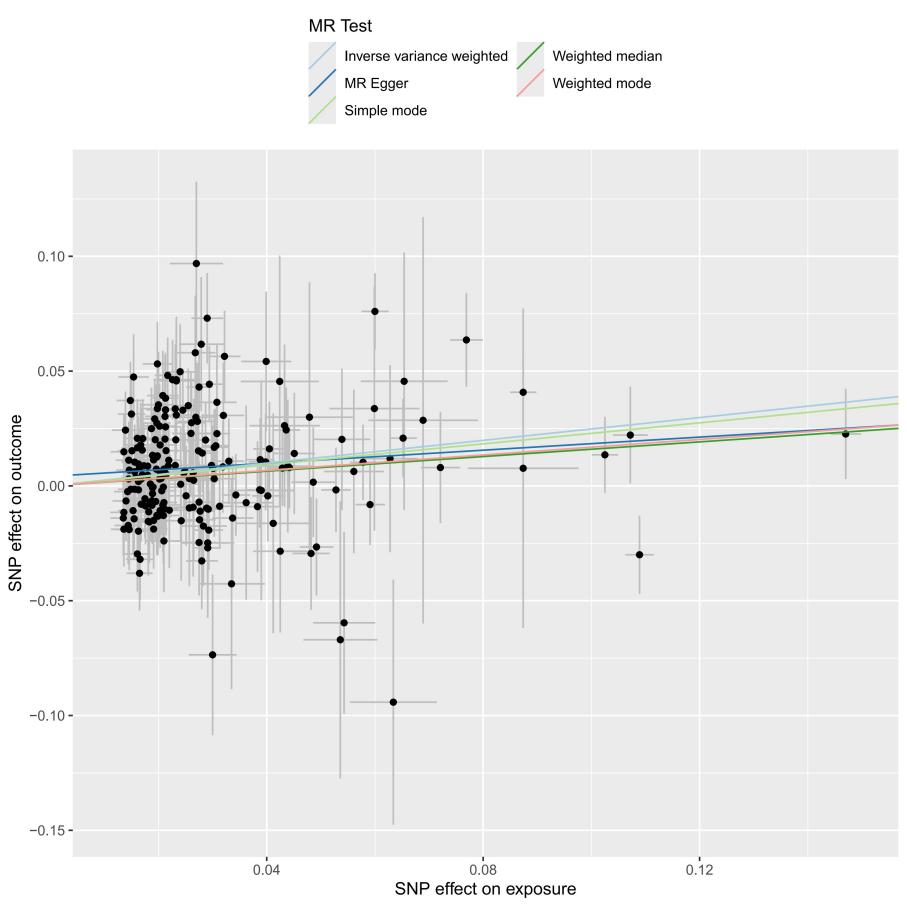
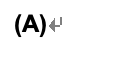
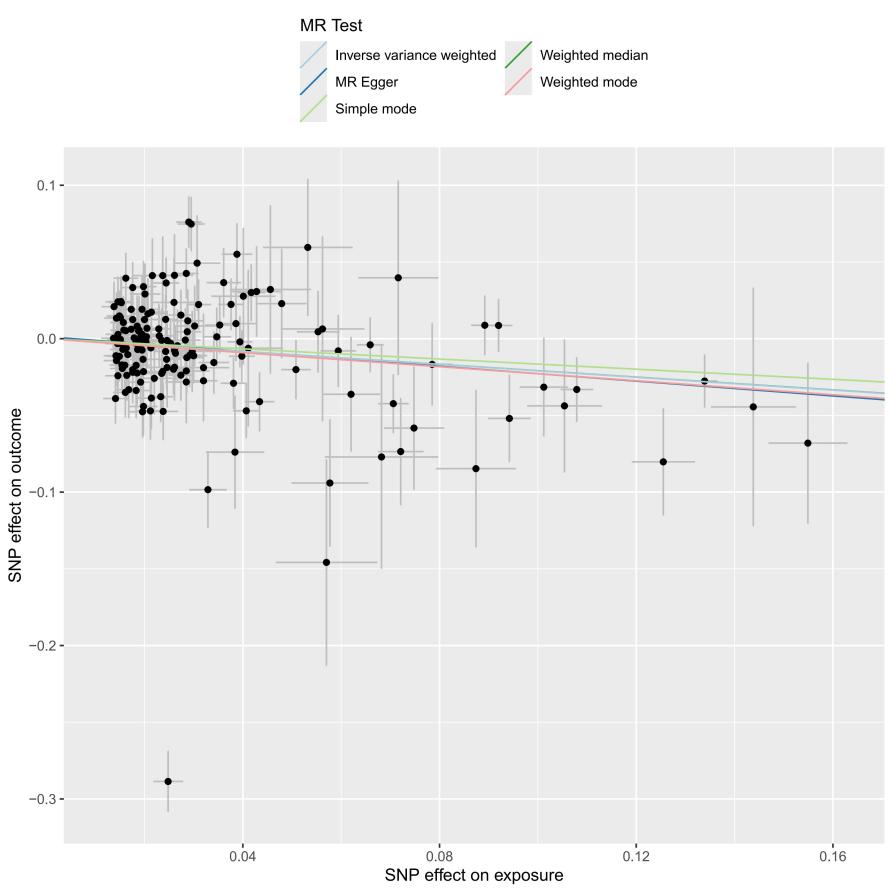
**(B)**

**(A)**



1. scatterplot of the MR analysis; (B) sensitivity analysis

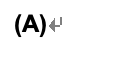
# Figure S6. Scatterplots of MR analysis between Apolipoprotein A (exposure), GGT (exposure), and MASLD (outcome).



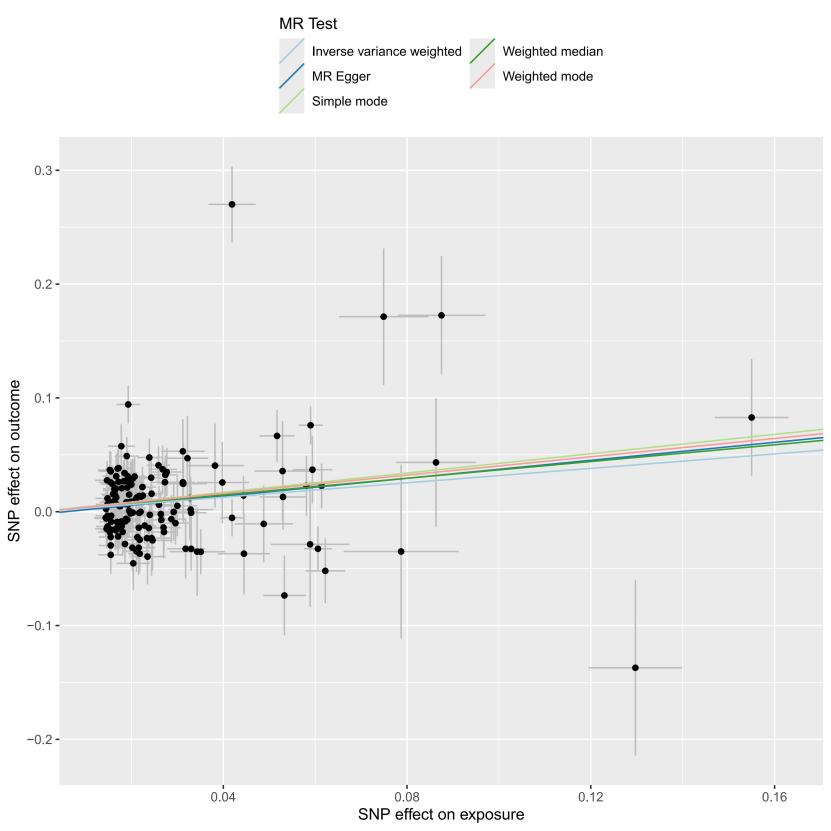
**(B)**

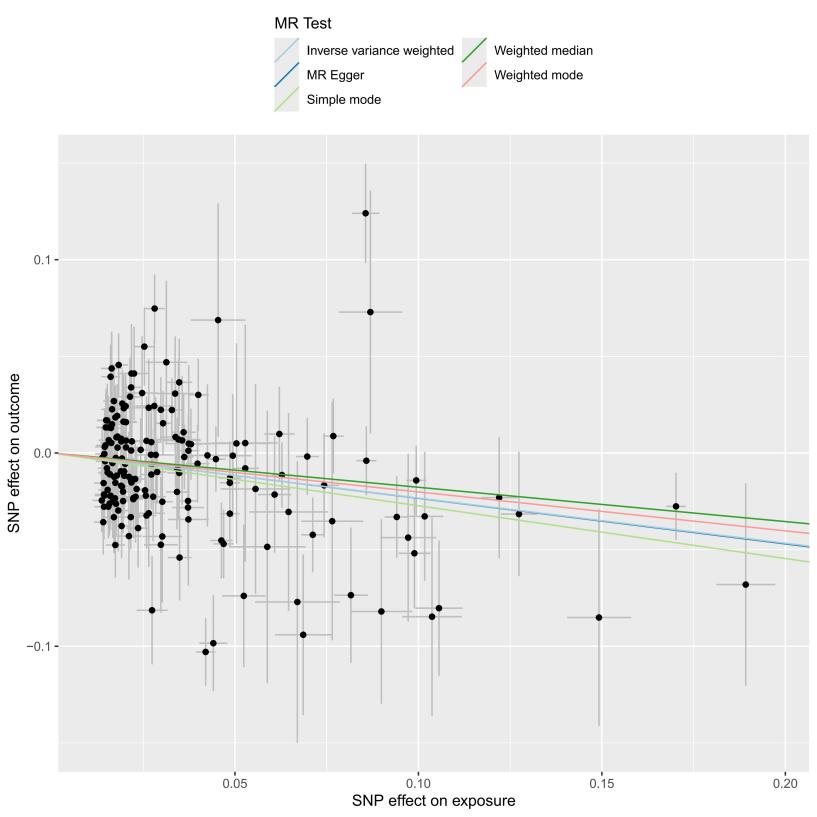
1. Apolipoprotein A; (B) gamma glutamyltransferase

# Figure S7. Scatterplots of MR analysis between albumin (exposure), HDL-C (exposure), and MASLD (outcome).



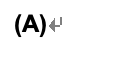
**(B)**



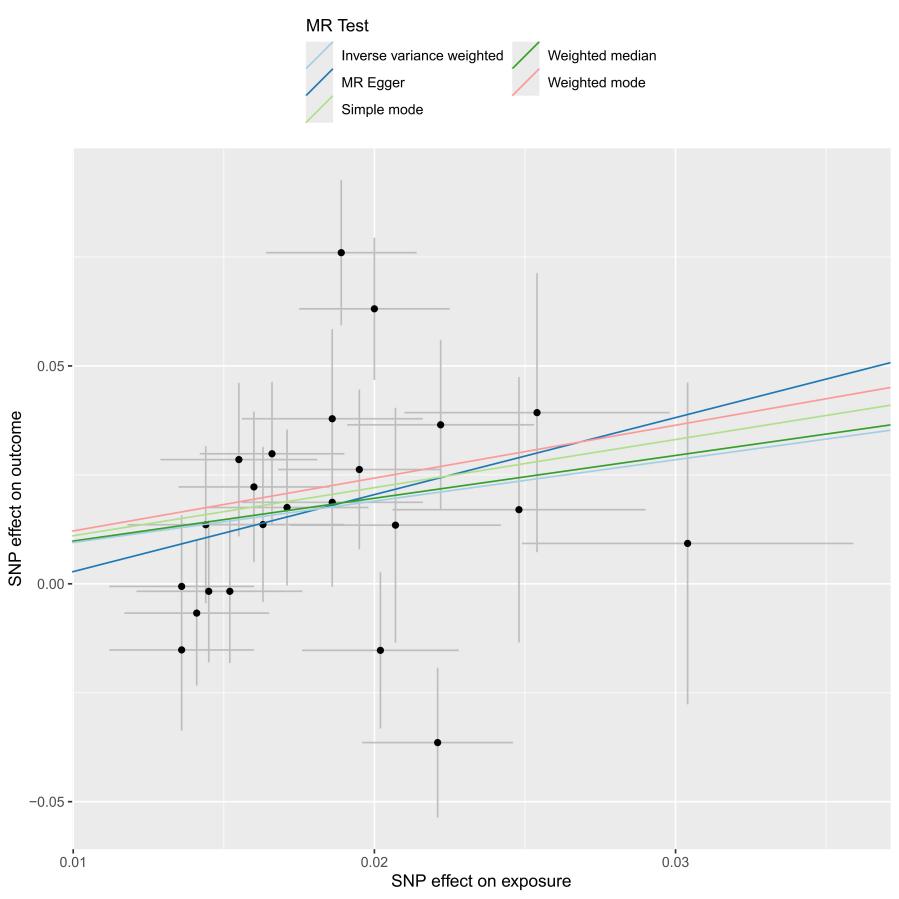
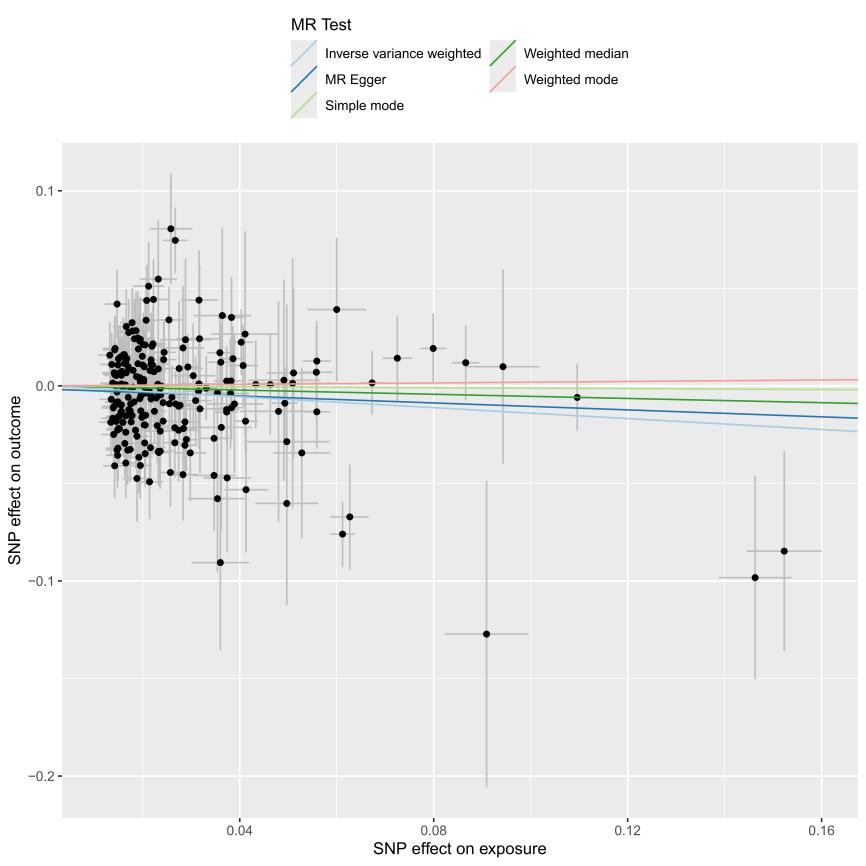


1. albumin; (B) HDL-C

# Figure S8. Scatterplots of MR analysis between IGF-1 (exposure), urinary sodium (exposure) and MASLD (outcome).

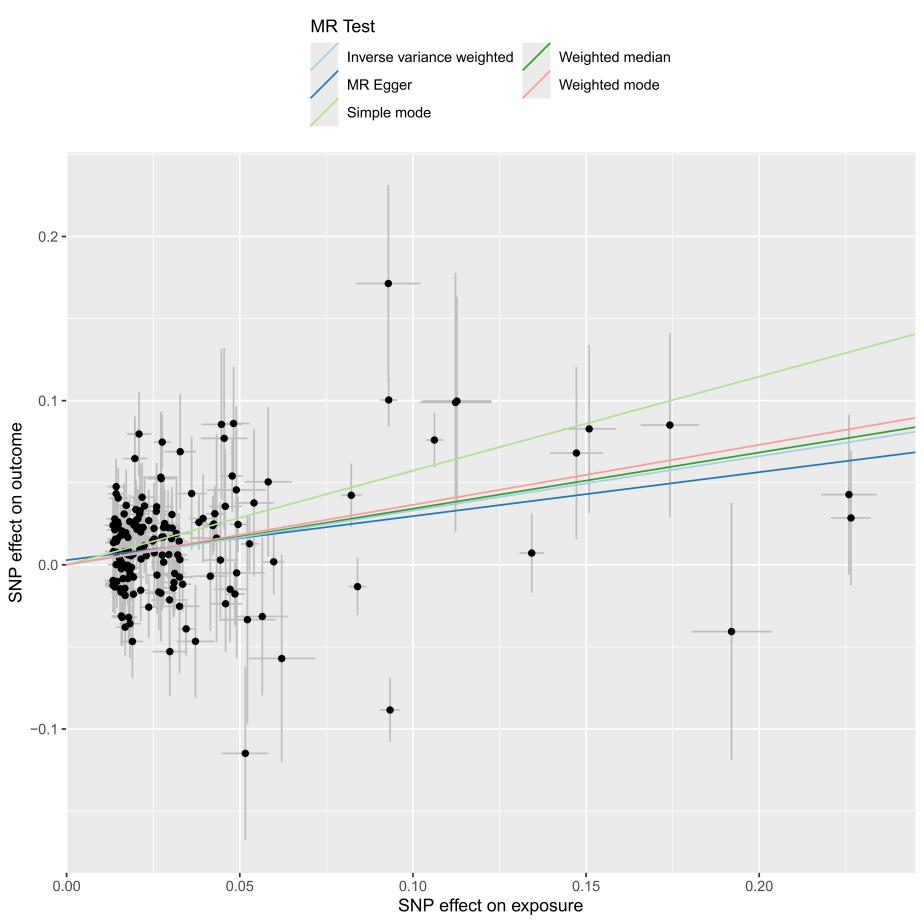


**(B)**

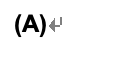
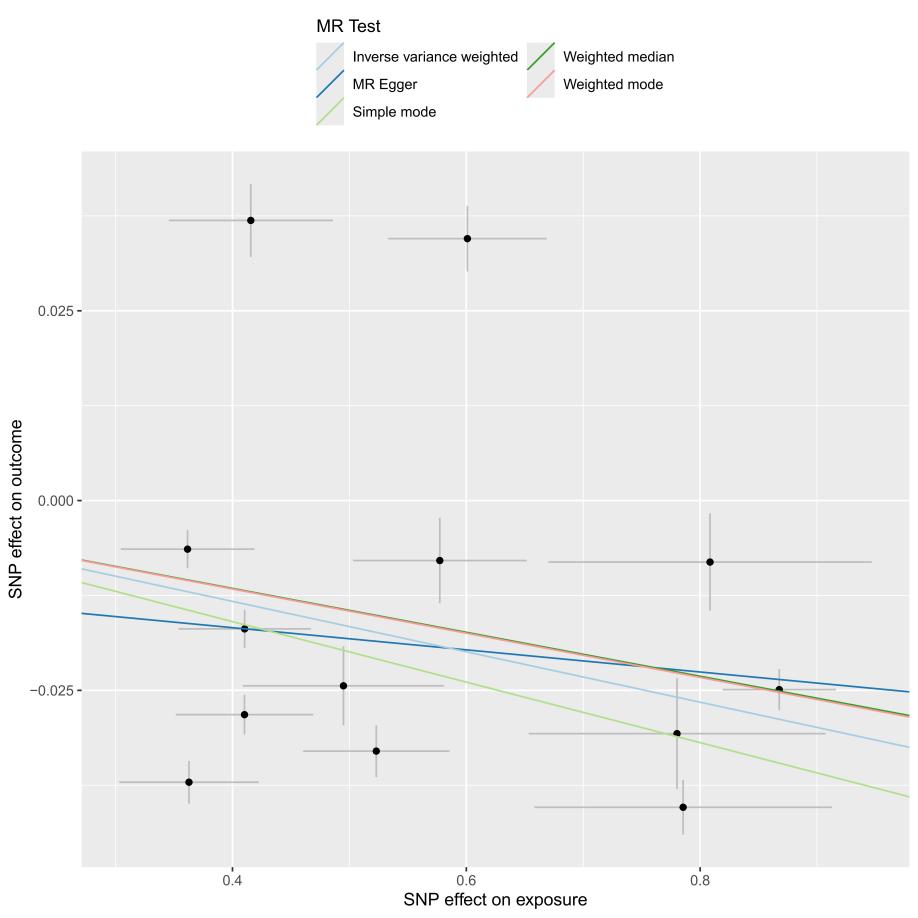


1. IGF-1; (B) urinary sodium

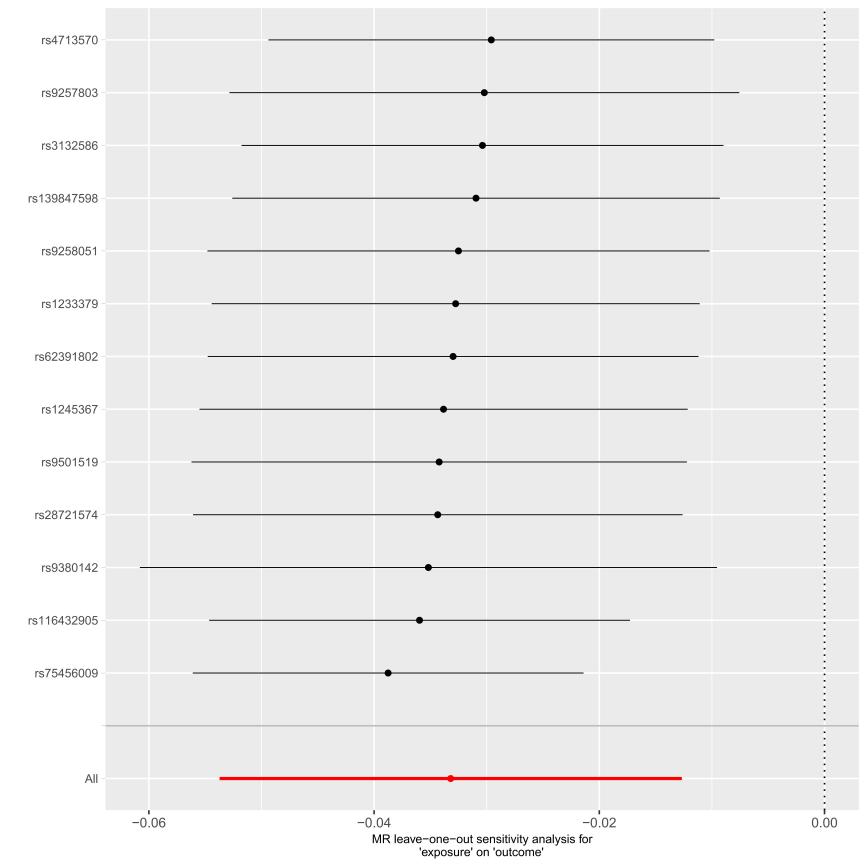
# Figure S9. Scatterplots of MR analysis between triglycerides (exposure) and MASLD (outcome).



# Figure S10. Causal relationship between HLA-A (exposure) and serum total protein levels (outcome).

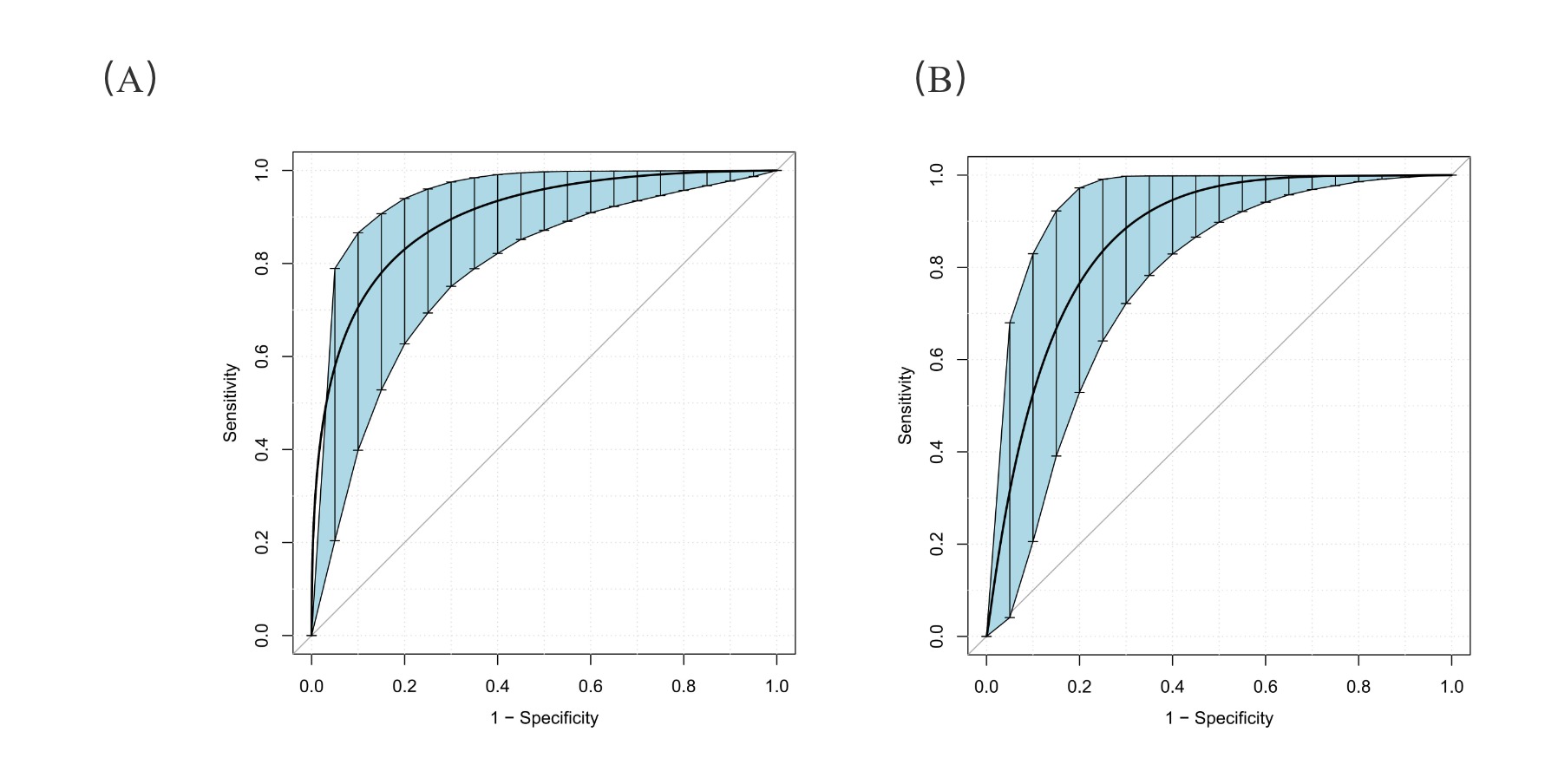


**(B)**



1. scatterplot of the MR analysis; (B) sensitivity analysis

# Figure S11. Bootstrap ROC curves for the random forest model in the training and validation sets



1. Bootstrap ROC curves for the random forest model in the training dataset; (B) Bootstrap ROC curves for the random forest model in the validation dataset.

# Figure S12. Differential expression analysis of CNPY4 in hepatocellular carcinoma and other cancer types.

图表, 箱线图

描述已自动生成(A)

图表, 雷达图

描述已自动生成(B)

1. Boxplot showing the expression levels of CNPY4 (log₂[TPM + 1]) in tumor (red) and normal (blue) tissues across various cancer types; (2) Radar plot comparing CNPY4 expression levels between tumor and normal tissues. The red line represents tumor tissues, and the blue line represents normal tissues.

*Note:* LIHC, liver hepatocellular carcinoma; TPM, transcripts per million.

# Figure S13. Association between CNPY4 expression and overall survival in hepatocellular carcinoma and other cancer types.

图片包含 图表

描述已自动生成

*Note:* LIHC, liver hepatocellular carcinoma.

# Figure S14. Correlation between CNPY4 expression and immune cell infiltration in hepatocellular carcinoma and other cancers.

图表, 散点图

描述已自动生成

*Note:* LIHC, liver hepatocellular carcinoma.

# Figure S15. Differential expression analysis of ENTPD6 in hepatocellular carcinoma and other cancer types.

(A)



(B)



1. Boxplot showing the expression levels of ENTPD6 (log₂[TPM + 1]) in tumor (red) and normal (blue) tissues across various cancer types; (B) Radar plot comparing ENTPD6 expression levels between tumor and normal tissues. The red line represents tumor tissues, and the blue line represents normal tissues.

*Note:* LIHC, liver hepatocellular carcinoma; TPM, transcripts per million.

# Figure S16. Association between ENTPD6 expression and overall survival in hepatocellular carcinoma and other cancer types.

图片包含 图表

描述已自动生成

*Note:* LIHC, liver hepatocellular carcinoma.

# Figure S17. Correlation between ENTPD6 expression and immune cell infiltration in hepatocellular carcinoma and other cancers.

图表, 散点图

描述已自动生成

*Note:* LIHC, liver hepatocellular carcinoma.