Optimizing Mendelian Randomization for Drug Prediction: Exploring Validity and Research Strategies

Miaoran Zhang

mrzhang20@mails.jlu.edu.cn

Jilin University https://orcid.org/0000-0001-6407-7750

Zhihao Xie

Jilin University

Aowen Tian

Jilin University

Zhiguo Su

Jilin University

Wenxuan Wang

Jilin University

Baiyu Qi

Jilin University

Jianli Yang

Jilin University

Jianping Wen

Jilin University

Peng Chen

Jilin University https://orcid.org/0000-0002-1422-4641

Research Article

Keywords: Mendelian randomization, Drug research, eQTL, pQTL, Linkage disequilibrium, Statistical models, Inflammatory Bowel Disease

Posted Date: February 28th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3966011/v1

Abstract

Mendelian randomization (MR) plays an increasingly important role in drug discovery, yet its full potential and optimized framework for accurately predicting drug targets have not been firmly established. This study aimed to evaluate the efficacy of multiple MR models in predicting effective drug targets and to propose the optimal selection of models and instrumental variables for MR analyses. We meticulously constructed datasets using approved drug indications and a range of IVs, encompassing cis-expression quantitative trait loci (eQTLs) and protein quantitative trait loci (pQTLs). Our analytical approach incorporated diverse models, including Wald's ratio, inverse-variance weighted (IVW), MR—Egger, weighted median, and MRPRESSO, to evaluate MR's validity in drug target identification. The findings highlight MR efficacy, demonstrating approximately 70% accuracy in predicting effective drug targets. For the selection of instrumental variables, tissue-specific eQTLs in disease-related tissues emerged as superior IVs. We identified a r^2 threshold below 0.3 as optimal for excluding redundant SNPs. To optimize the MR model, we recommend IVW as the primary computational model, complemented by the weighted median and MRPRESSO for robust analyses. This finding is consistent with current findings in the literature. Notably, a P value of < 0.05, without false discovery rate correction, is the most effective for identifying significant drug targets. With the optimal strategies we summarized, we identified new potential therapeutic targets for IBD and its subtypes, including ERAP1, HLA-DQA1, IRF5 and other genes. This study provides a refined, optimized strategy for MR application in drug discovery. Our insights into the selection of instrumental variables, model preferences, and parameter thresholds significantly enhance MR's predictive capacity, offering a comprehensive guide for future drug development research.

Introduction

The journey of drug development is a challenging and resource-intensive venture. It involves substantial investment in terms of time, labor, and financial resources, yet the efficiency of research and development in this domain has been declining. [1, 2],. Recently, Mendelian randomization (MR) has surfaced as a promising tool for enhancing the efficiency of drug development. This method leverages instrumental variables (IVs) to predict causal relationships between potential drug targets and diseases, offering a robust approach for both discovering novel drugs and repurposing existing drugs [3]. For instance, studies such as those by Solal Chauquet et al. and Jie Zheng et al. have demonstrated the utility of MR in identifying new associations between known drug targets and diseases, including psychiatric disorders and a wide range of phenotypes [4, 5]. Additionally, MR-based findings have been instrumental in suggesting potential drug targets for diseases such as Parkinson's disease [6].

Despite these advancements, the efficacy of MR in drug target discovery is not entirely clear. Many of the targets identified through MR have not yet been experimentally validated, leaving their practical effectiveness in question. Key factors influencing the detection of targets in MR studies include the choice of instrumental variables, such as expression quantitative trait loci (eQTLs) and protein quantitative trait loci (pQTLs), and the debate over the best type of QTL for reliable drug ability assessments. The lack of consensus on the optimal correlation coefficient for removing linkage

disequilibrium among SNPs and the appropriate threshold for defining significant results further add to the complexity. Moreover, variations in modeling approaches for estimating causal effects indicate the need for a more standardized and precise methodology in MR analyses [7] [8].

To address these challenges, our study undertook a comprehensive analysis by organizing datasets of all approved drugs that have passed phase III clinical trials, along with their target proteins. We constructed both true-positive and true-negative datasets of drug-target-indication pairs to conduct extensive MR studies. Our objective was to delineate the effectiveness of MR in drug discovery and provide strategic parameter guidance for future drug development.

Methods

Target protein identification

For the pivotal step of identifying target proteins, our approach began with extracting approved drug indications, particularly drug-indication combinations, from the ChEMBL database (version 29). ChEMBL is a comprehensive database cataloging bioactive molecules. It integrates extensive data on drug molecules, bioactivity, and genomics, thereby facilitating the development of new drugs with a genomic foundation.

Next, we systematically identified and selected drugs associated with these approved indications. We then utilized the ChEMBL API (see Supplementary Materials) to obtain the corresponding UniProt IDs of the target proteins. In instances where a drug was linked to multiple protein targets, we included each of these targets in our analysis to ensure a comprehensive evaluation. This methodological rigor allowed us to construct an inclusive dataset that was critical for the subsequent steps of our MR analysis.

GWAS summary data acquisition

Our approach to obtaining genome-wide association studies (GWAS) summary data for indications involved a consensus-driven process. This process was described by a review panel comprising three independent research participants. Each participant was subjected to a thorough search of the Integrative Epidemiology Unit (IEU) GWAS database (https://gwas.mrcieu.ac.uk/) to identify pertinent GWAS data aligned with the indications relevant to our study. To ensure robustness in our selection, we adopted a data validation strategy: if two or more participants agreed on the same GWAS dataset, it was directly incorporated into our analysis. In scenarios where there was a divergence of opinions, the panel engaged in detailed discussions to arrive at a consensus and make a final determination.

To broaden the scope and applicability of our study, we implemented strategic replacements for some GWAS datasets that were unavailable or incomplete. Our replacement methodologies included the following:

1. The disease subtypes were substituted for the overarching disease data when specific subtype data were absent.

2. Data from a representative subtype were used in place of the entire disease dataset.
3. Interchanging data between closely related subtypes of the same disease.
4. GWAS data are exchanged between diseases known to have causal relationships.

This meticulous process ensured comprehensive coverage and relevance of GWAS data, thereby augmenting the validity of our Mendelian randomization analysis.

Types of instrumental variables

The IVs employed in our MR analysis can be primarily divided into three distinct types:

Plasma pQTLs

We sourced plasma protein pQTL data from the research conducted by Benjamin B. Sun et al. [9]. This study comprehensively measured 2994 plasma proteins across 3301 participants of European ancestry using the advanced SomaLogic SomaScan platform. The data from this research are publicly available and provide a robust foundation for our plasma protein-focused analyses.

Tissue-specific eQTLs

Utilizing the resources of the Gene-Tissue Expression project (GTEx V.8), we obtained eQTL data for genes across 48 different tissues, including blood. This extensive dataset allowed us to incorporate a broad range of tissue-specific eQTLs as instrumental variables in our MR studies, enhancing the diversity and applicability of our analysis.

Brain pQTLs

We accessed brain tissue pQTL data from a study by Chloe Robins et al., which focused on genetic and proteomic analyses of 330 postmortem samples from the lateral prefrontal cortex of elderly individuals [10]. This study reported pQTL data for 7376 brain proteins, providing us with a comprehensive dataset for brain-specific MR analysis.

More information about the GWAS summary data of the target proteins can be found in Supplementary Table 1–3.

Composition of IV-Indication Pairs

To evaluate the impact of different frameworks on drug MR analysis, we generated sets of IV-indication pairs classified as either “true positive” or “true negative”.

True-positive target-indication pairs

These pairs were derived from approved drug indications, with drug targets as the exposure and the indications as the outcome. Based on the instrumental variables used, we constructed five distinct sets of true positive pairs. These include plasma protein pQTLs for all related indications, plasma pQTLs for

blood-related indications, blood eQTLs for blood-related indications, tissue-specific eQTLs for all related indications, and brain pQTLs for brain indications. The underlying assumption for these pairs is the existence of a causal relationship between the drug targets and indications, as evidenced by clinical trial efficacy, thus suggesting potential positive MR results. This method allowed us to assess the true positive rate and false negative rate under various conditions.

True-negative target-indication pairs

These pairs were used to evaluate the ability to correctly identify negatives. We merged eQTLs from different tissues with plasma pQTLs to create a comprehensive SNP pool. For each indication in the true-positive target-indication pairs, the same number of SNPs were randomly sampled from the SNP pool. Since these pairs are constructed with hypothetical target proteins without any known causal relationship to the indications, the anticipated MR results for these pairs are negative.

Determining the optimal LD r^2 threshold

In MR analysis, removing redundant SNPs by LD serves as a balance between the bias induced by correlated SNPs and the instrumental power. To find an optimal LD r^2 threshold, we applied various r^2 values (0.001, 0.01, 0.2, 0.3, 0.4, 0.5, and 0.6) within a 500 kb clumping window to remove LD in the tissue-specific eQTLs of *true-positive IV-indication pairs*. Next, we performed MR analysis across different models to observe the effect of varying LD r^2 values on the true positive rate. The optimal LD r^2 value was ascertained based on these analyses. This value represents the threshold at which the true positive rate is maximized while avoiding the introduction of bias due to excessive LD among the IVs.

Determining the optimal P value threshold

Next, we clumped the instrument variables in the positive datasets of tissue-specific eQTLs using the optimal LD r^2 obtained in the previous steps and performed MR analysis in both the positive and negative datasets. FDR correction was then applied to the P values in the obtained results. Subsequently, we screened each model's original and FDR-corrected P values using four commonly used filtering thresholds of 0.1, 0.05, 0.01, and 0.001 to obtain significant results and constructed a confusion matrix of each model under each P value threshold to determine the optimal P value threshold.

Effect of instrumental variables and model selection on the effectiveness of drug MR

We utilized the entire dataset, including the true positive and true negative datasets, to calculate causal relationships between exposure and outcome using five different algorithms. Subsequently, we constructed confusion matrices to evaluate the performance across different datasets and determine which algorithm model exhibited the highest true positive rate. This analysis aimed to assess the most effective dataset and identify the algorithm with the highest true positive rate.

Mendelian randomization analysis

Using the R packages TwoSampleMR (version 0.5.6) and MRPRESSO (version 1.0), we conducted MR analysis by applying five models: the Wald ratio, inverse variance weighted (IVW), MR–Egger, weighted median, and MRPRESSO models. When heterogeneity existed, we used the results from the IVW random effects model as the final effect size. Otherwise, we used the fixed effect model. To ensure the validity of MR analysis, the selected instrumental variables (IVs) should meet the following three critical principles: (1) the IVs should be strongly associated with the gene expression level or protein expression level ($P < 5\times10^{-8}$). (2) There should be no association between the IVs and the outcome variables (i.e., the indications). (3) There should be no association between the IVs and confounding factors.

In this study, $\widehat{\beta}_j$ represents the causal effect of exposure on the outcome, $\widehat{\gamma}_j$ represents the strength of the association between genetic variants and exposure, and $\widehat{\gamma}_j$ is the regression coefficient of the outcome on each genetic variant. For cases with multiple instrumental variables, we employed five models for MR calculation: the Wald ratio, IVW, MR–Egger, weighted median, and MRPRESSO models[11–13].

Statistical analysis

We defined a drug indication (i.e., a combination of a disease and a drug) as a sample and performed statistics on the results of five models: Wald ratio, IVW, MR–Egger, weighted median, and MRPRESSO. The total sample size N for each model is the total number of drug indications in the results. Then, we screened significant results using a selected P value threshold and excluded results that were driven by pleiotropy. In the true positive datasets, the direction of the effect of the target proteins on the significant results should also be the same as that reported in the ChEMBL database. For drug indications composed of multitarget drugs, we considered the sample's result to be significant if the result of any target was significant. We then counted the number of significant samples for each model and obtained the positive sample size n for the model. The true positive rate and false positive rate of drug prediction using the model can be calculated from the formula n/N in the positive and negative datasets, respectively.

In MR analysis, we used Cochrane's Q test to assess heterogeneity and MR–Egger regression to assess pleiotropy. Then, we calculate the F statistic, which measures the effectiveness of the instrumental variables[14]. It is important to note that the method used to calculate r^2 is only an approximation due to the lack of sample genotype data.

In this study, we used positive and negative datasets under the same conditions and calculated the following metrics:

The accuracy represents the proportion of correct predictions out of the total samples and is expressed as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision refers to the probability of a positive sample actually being positive, given that it has been predicted to be positive. It is expressed as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall is a metric that considers the original sample. Its meaning is the probability of an actual positive sample being predicted as positive. The expression is as follows:

$$\text{Recall} = \frac{TP}{TP + FN}$$

The F1-score is the harmonic mean of precision and recall, which considers both precision and recall to reach the highest level simultaneously and achieve a balance.

$$\text{F1} - \text{Score} = \frac{2 * \text{Precision*Recall}}{\text{Precision} + \text{Recall}}$$

Results

Study design

We initiated our study by identifying drugs beyond phase III of clinical trials and associating drugs with their target proteins. In the pursuit of optimal thresholds for removing linkage disequilibrium and significance levels, we employed tissue-specific expression quantitative trait loci of target proteins as instrumental variables. Subsequently, we systematically evaluated the efficacy of Mendelian randomization predictions using pQTLs and eQTLs across five commonly used models, probing the impact of IV selection on outcomes (Fig. 1). Finally, to further validate the superiority of our MR strategy, we applied the optimal strategy to identify potential drug targets for inflammatory bowel disease.

Drug indications

We collected a dataset encompassing 4,309 pairs of approved drug indications representing distinct drug-indication combinations sourced from the ChEMBL database (V.29). For each approved indication, we meticulously extracted the corresponding drugs and obtained the UniProt IDs of their target proteins using the ChEMBL API. (Supplementary Table 1–1). Additionally, in accordance with the outlined methodologies, we assessed genome-wide association study (GWAS) summary statistics for 265 diseases (Supplementary Table 1–2).

Optimization of the LD r^2 threshold for IV selection

In MR analysis, removing redundant SNPs by LD serves as a balance between the bias induced by correlated SNPs and the instrumental power. To find an optimal LD cutoff, we applied various r^2 values (0.001, 0.01, 0.2, 0.3, 0.4, 0.5, and 0.6) within a 500 kb clumping window to remove LD in the tissue-

specific eQTLs of *true-positive target-indication pairs*, followed by performing MR analysis using the IVs obtained under these varying conditions.

Our analysis, as illustrated in Fig. 2, revealed key insights into the impact of different LD r^2 thresholds. We noted that a r^2 value greater than 0.01 induced a gradual increase in the true positive rate across all models (IVW, MR–Egger, weighted median, MRPRESSO), except for the Wald ratio. The true positive rate initially increased at $r^2 = 0.2$, with the growth rate decelerating beyond $r^2 = 0.4$ and eventually stabilizing. A significant observation was the plateauing of the MR–Egger model at a r^2 of 0.3.

This led to the conclusion that an LD r^2 threshold of 0.3 is optimal for minimizing the impact of LD, as evidenced by consistently high true positive rates surpassing 0.7 in the IVW, weighted median, and MRPRESSO models. Beyond this threshold, while the true positive rate exhibited further increases, the F-statistic of the IVs showed no significant change (violin plot in Fig. 2). This indicates that for a r^2 value higher than 0.3, despite increasing the number of IVs, there is no significant improvement in the F statistics and that the additional instrumental variables (IVs) may lack effectiveness. This lack of effectiveness could be attributed to the presence of greater linkage disequilibrium and undetectable SNP-level pleiotropy, resulting in a biased estimation of the effect size of MR. In summary, our comprehensive evaluation revealed that a r^2 threshold of 0.3 optimally balances the need to minimize LD while maintaining the integrity and accuracy of MR predictions. Complete results on true positive rates and SNP counts across various LD r^2 thresholds are provided in Supplementary Table 2.

Establishing the optimal P value threshold for positive MR results

A pivotal aspect of MR studies is to determine the most effective P value threshold for selecting significant MR results. We examined the true positive rate and precision of using different P value cutoffs in the selection of positive MR results. The significant tissue-specific eQTLs in the true-positive target-indication pairs were clumped using the previously established optimal LD r^2 value (0.3). MR analysis was then conducted for both the true-positive and true-negative datasets.

To refine our results, we applied false discovery rate (FDR) correction to the MR P values obtained from these analyses. Subsequently, we embarked on a detailed evaluation of various commonly used P value filtering thresholds. This assessment encompassed both raw P values and FDR-corrected P values, with a focus on the true positive rate and precision as key metrics.

Our nuanced analysis, depicted in Fig. 3, revealed distinct patterns within the correct and incorrect P value groups. We noted a decrease in the true positive rate for both the adjusted and unadjusted groups as the P value threshold decreased. The difference in TPRs between the P values of 0.1 and 0.05 was not statistically significant. However, setting the P value at 0.05 yielded greater precision than setting the P value at 0.1. Based on these findings, we advocate for a raw P value threshold of $P < 0.05$. (Supplementary Table 3)

Impact of Instrumental Variables on MR Analysis Efficacy

We further examined the effect of different types of IVs and the choice of computational models on the efficacy of drug MR analysis.

We conducted MR analyses across all the true positive datasets and true negative datasets using five commonly used models, focusing on the TPR, TNR, accuracy and precision of each model. Our results clearly showed the superior effectiveness of tissue-specific pQTLs and eQTLs as IVs. Datasets utilizing plasma pQTLs in blood-related diseases and tissue-specific eQTLs in corresponding tissue-related diseases demonstrated higher performance, with accuracies consistently above 0.7 and a TPR of nearly 0.9. Specifically, blood tissue eQTLs related to blood-related diseases achieved remarkably high scores in the IVW model, with the accuracy and TPR reaching 0.87 and 0.98, respectively.

Conversely, the use of brain tissue pQTLs in brain-related diseases yielded less favorable results, with lower accuracy and TPR across all models.

Moreover, our analysis highlighted the inverse-variance weighted (IVW) model as the most effective model for MR analysis in drug target prediction. The IVW model consistently outperformed the other models, especially in datasets with tissue-specific eQTLs, showing the highest TPR. Acknowledging the unique strengths of each model, we advocate the weighted median model as a valuable complement to the IVW model. This is evidenced by the highest overlap of true positive results between them, with 1038 and 803 in the positive and negative datasets, respectively. Intriguingly, the MR—Egger model, while having a notably low false positive rate of less than 5% in each dataset, exhibited the lowest overlap with the IVW model except for the wald ratio, with 272 and 149 overlaps in the positive and negative datasets, respectively.

These insights into the selection of IVs and computational models significantly contribute to optimizing the efficacy of MR in predicting drug targets (Fig. 4, Supplementary Table 4).

Novel Therapeutic Targets for Inflammatory Bowel Diseases

To further validate the superiority of our MR strategy, we conducted a comprehensive analysis aiming to uncover potential drug targets for IBD and its subtypes (each gene and its corresponding drug information are listed in detail in Supplementary Table 5–12).

After a comprehensive search of the ChEMBL database, we identified 157 therapeutic drugs for IBD and CD but none for UC. Using the optimized MR strategy mentioned above, we verified that 58 of the 157 drugs were effective for treating IBD and CD, with 7 specific for treating IBD and 51 for treating CD. Additionally, for the targets of these drugs, our analysis revealed 11 target genes for IBD and 33 target genes for CD (Supplementary Table 13, Supplementary Fig. 1).

Interestingly, we also discovered numerous novel causal genes that serve as drug targets for other diseases. Among all the novel genes, 24 were unique to 36 genes, indicating that the genes showed causal relationships with both IBD and its subtypes, and these genes were found in all tissues (Fig. 5A).

Pathway analysis revealed that these genes are predominantly involved in immune response pathways (Fig. 5B). Notably, genes such as IRF5 and HLA-DQA1 had high odds ratios (ORs) in multiple tissues, suggesting that these genes play a significant role in IBD pathogenesis. ERAP1 and ERAP2 may have some therapeutic effect bias. ERAP2 consistently demonstrated a risk effect across all tissues in IBD and CD patients but showed no causal relationship in some tissues of UC patients. Conversely, ERAP1 exhibited a risk effect in various UC tissues but lacked strong consistency across tissues in IBD and CD patients (Fig. 5C). Furthermore, mining the ChEMBL database revealed that the protein encoded by TUFM may serve as a target for tyrosine kinase inhibitors (TKIs), such as dasatinib and lenvatinib, which play a role in the treatment of some cancers (Fig. 5D).

Discussion

In this study, for the first time, a comprehensive evaluation of MR in drug discovery was conducted, and MR demonstrated an effectiveness of up to 70%. Specific recommendations for several key parameters were made: a r^2 value of 0.3 is suggested for LD, a P value threshold of 0.05 without FDR correction is recommended, the selection of tissue-specific eQTLs or pQTLs as IVs is advocated, and the use of IVW as the primary calculation model, supplemented by the weighted median model, is proposed. Through an in-depth study of the drug MR, we provide a powerful strategy for evaluating and optimizing MR parameters and model selection, which can further enhance the efficiency of MR in drug development and explore more potential drug targets.

Specifically, our results show that with a r^2 value of 0.3 and a P value threshold of 0.05 without FDR correction, the predictive accuracy and recall of MR are optimized. In MR studies, overly strict LD removal standards can result in too few available instrumental variables, leading to a decrease in statistical power, while overly loose standards can introduce bias due to horizontal pleiotropy[8]. Our recommended parameter settings can effectively balance the ratio of false positives and false negatives [15] as well as horizontal pleiotropy and statistical power, ensuring prediction accuracy while discovering as many effective drug targets as possible. Although performing FDR correction can effectively reduce the false positive rate of predictions[16] and has been adopted by many existing drug MR studies[17], it was not as effective as the researchers thought. As our results showed, its effect on TPR is very strong and can lead to the omission of many potential drug targets in our study, which contradicts our objective.

Furthermore, we found that using tissue-specific eQTLs or pQTLs as IVs, with IVW as the primary computational model supplemented by the weighted median model, can further enhance the precision of MR, and they have been widely adopted by the majority of researchers[18, 19]. For instance, we found that using tissue-specific eQTLs or pQTLs as IVs was very effective in improving MR predictive accuracy, consistent with the findings of Liam Gaziano et al. [7]. They identified six potential treatment targets for

COVID-19 using tissue-specific eQTLs from 49 different tissues as IVs, including several reported genes such as ACE2[20, 21].

To further validate our optimal study strategy, we focused on IBD and its subtypes in drug MR studies. Despite the increase in medical therapies for IBD, including targeted drugs such as TNF, IL-23, JAK, and phosphodiesterase inhibitors, their effectiveness remains limited [22, 23]. This underscores the need for novel targeted drugs for IBD treatment. Our study highlights the association of IBD with MHC class I and ERAP1 interactions, revealing a common immunopathogenic foundation[24]. We confirmed the uniform risk effect of ERAP1 across IBD subtypes, suggesting its potential as a therapeutic target. Additionally, high ORs for genes such as IRF5 and HLA-DQA1 in multiple tissues indicate their significant role in IBD pathogenesis. Notably, the HLA-DQA1*05 allele, which is implicated in predisposing patients to ulcerative colitis and is associated with the development of antibodies against TNF antagonists such as infliximab, is linked to the immunogenicity of anti-TNF agents, particularly in the presence of the HLA-DQA1*05 allele group [25]. Furthermore, combining our analysis of the ChEMBL database, we identified TUFM as a potential target among 25 novel targets. TUFM may serve as a tyrosine kinase inhibitor target. These inhibitors exert therapeutic effects on certain types of cancer by suppressing the activity of tyrosine kinases, thereby influencing cellular signaling pathways. Consequently, they are the key components of numerous cancer-targeting drugs [26]. The use of tyrosine kinase inhibitors is associated with a significant number of side effects, possibly due to their targeting of not only tyrosine kinases but also other targets. Current literature primarily reports on their impact on cardiovascular events[27], and our research provides a cautionary note regarding the potential adverse effects of current cancer-targeting drugs on gastrointestinal diseases

Limitations

Our study has several limitations. Due to the complex and unclear pathogenesis of many diseases, it is difficult to accurately determine the tissues associated with specific diseases. Moreover, the random sampling method we used to construct the negative queue may increase the false-negative rate to some extent when the number of IVs is very small. Despite these limitations, our research results still have strong practicality and application potential, providing important guidance for drug development.

Conclusions

In summary, our study systematically evaluated the MR approach for predicting drug targets utilizing positive and negative datasets based on existing drugs, targets, and indications. We found that the MR method achieved an effective accuracy of up to 70% in predicting drug targets. Optimal parameters were identified, including an LD r^2 of 0.3 for removing linkage disequilibrium in instrumental variables and a significance threshold of the original P value < 0.05 for result significance. Tissue-specific eQTLs or pQTLs in disease-related tissues emerged as the most effective instrumental variables. We recommend the IVW model as the primary choice, complemented by the weighted median model. We also applied this

strategy to explore potential therapeutic targets for IBD and its subtypes, providing evidence for candidate genes such as IRF5, ERAP1, and HLA-DQB1.

Abbreviations

MR - Mendelian randomization

IVs - Instrumental Variables

eQTLs - Expression Quantitative Trait Loci

pQTLs - Protein Quantitative Trait Loci

IVW - Inverse-variance weighted

FDR - False discovery rate

GWAS - Genome-Wide Association Studies

IBD - Inflammatory Bowel Disease

CD - Crohn's Disease

LD - Linkage Disequilibrium

SNPs - Single nucleotide polymorphisms

OR - odds ratio

TNF-Tumor Necrosis Factor

ADA - Anti-Drug Antibodies

TP - true positive

TN - true negative

FP - False positive

FN - False negative

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The data used in this paper are from public databases as following, IEU database(https://gwas.mrcieu.ac.uk/), GTEX-v8 (https://www.gtexportal.org/home/). The dataset(s) supporting the conclusions of this article is(are) included within the article (and its additional file(s)).

Competing interests

The authors declare that there are no conflicts of interest.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the “Changbai Mountain” distinguished professor project of Jilin Provence, China.

Authors' contributions

Miaoran Zhang and Zhihao Xie played major roles in the analysis of the data and the drafting of the manuscript. Aawen Tian participated in the analysis of the data and revision of the manuscript. Zhiguo Su and Wenchuan Wang participated in the analysis of the data. Baiyu Qi, Jianli Wang, and Jianping Wen participated in the study design. Peng Chen played a major role in the study design and acquisition of the data. All the authors have read and approved the final manuscript.

Acknowledgments

The drug target data were obtained from the ChEMBL database. The plasma protein data were obtained from the INTERVAL cohort and the Icelanders cohort. The GWAS summary statistics were obtained from public data and formatted by the MRC-IEU. We are very grateful to them for publishing the results of this GWAS. We also thank the MRC Integrative Epidemiology Unit (IEU) at the University of Bristol for the manually curated collection of complete GWAS summary datasets.

References

1. Berdigaliyev N, Aljofan M. An overview of drug discovery and development. Future Med Chem. 2020;12(10):939–47.
2. Scannell JW, et al. Diagnosing the decline in pharmaceutical R&D efficiency. Nat Rev Drug Discov. 2012;11(3):191–200.
3. Walker VM, et al. Mendelian randomization: a novel approach for the prediction of adverse drug events and drug repurposing opportunities. Int J Epidemiol. 2017;46(6):2078–89.

4. Chauquet S, et al. Association of Antihypertensive Drug Target Genes With Psychiatric Disorders: A Mendelian Randomization Study. JAMA Psychiatry. 2021;78(6):623–31.
5. Zheng J, et al. Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. Nat Genet. 2020;52(10):1122–31.
6. Storm CS, et al. Finding genetically supported drug targets for Parkinson's disease using Mendelian randomization of the druggable genome. Nat Commun. 2021;12(1):7342.
7. Gaziano L, et al. Actionable druggable genome-wide Mendelian randomization identifies repurposing opportunities for COVID-19. Nat Med. 2021;27(4):668–76.
8. Schmidt AF, et al. Genetic drug target validation using Mendelian randomization. Nat Commun. 2020;11(1):3255.
9. Sun BB, et al. Genomic atlas of the human plasma proteome. Nature. 2018;558(7708):73–9.
10. Robins C, et al. Genetic control of the human brain proteome. Am J Hum Genet. 2021;108(3):400–10.
11. Hemani G et al. *The MR-Base platform supports systematic causal inference across the human phenome*. ELife, 2018. 7.
12. Burgess S, Thompson SG. Interpreting findings from Mendelian randomization using the MR–Egger method. Eur J Epidemiol. 2017;32(5):377–89.
13. Verbanck M, et al. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. Nat Genet. 2018;50(5):693–8.
14. Papadimitriou N, et al. Physical activity and risks of breast and colorectal cancer: a Mendelian randomization analysis. Nat Commun. 2020;11(1):597.
15. McMurray JJV et al. *Angiotensin-neprilysin inhibition versus enalapril in heart failure*. N Engl J Med, 2014. 371(11).
16. Shuken SR, McNerney MW. Costs and Benefits of Popular P Value Correction Methods in Three Models of Quantitative Omic Experiments. Anal Chem. 2023;95(5):2732–40.
17. Rasooly D, et al. Genome-wide association analysis and Mendelian randomization proteomics identify drug targets for heart failure. Nat Commun. 2023;14(1):3826.
18. Ding M, et al. Association between periodontitis and breast cancer: two-sample Mendelian randomization study. Clin Oral Investig. 2023;27(6):2843–9.
19. Yang M et al. *No Evidence of a Genetic Causal Relationship between Ankylosing Spondylitis and Gut Microbiota: A Two-Sample Mendelian Randomization Study*. Nutrients, 2023. 15(4).
20. Su W-L, et al. COVID-19 and the lungs: A review. J Infect Public Health. 2021;14(11):1708–14.
21. Mannar D et al. *SARS-CoV-2 Omicron variant: Antibody evasion and cryo-EM structure of spike protein-ACE2 complex*. Science (New York, N.Y.), 2022. 375(6582): p. 760–4.
22. Misselwitz B, et al. Emerging Treatment Options in Inflammatory Bowel Disease: Janus Kinases, Stem Cells, and More. Digestion. 2020;101(Suppl 1):69–82.
23. Singh S, et al. Systematic review with network meta-analysis: first- and second-line pharmacotherapy for moderate-severe ulcerative colitis. Volume 47. Alimentary pharmacology & therapeutics; 2018. pp.

162–75. 2.

24. McGonagle D, et al. MHC-I-opathy'-unified concept for spondyloarthritis and Behçet disease. Nat Rev Rheumatol. 2015;11(12):731–40.
25. Nowak JK et al. *HLA-DQA1*05 Associates with Extensive Ulcerative Colitis at Diagnosis: An Observational Study in Children*. Genes (Basel), 2021. 12(12).
26. Du Z, Lovly CM. Mechanisms of receptor tyrosine kinase activation in cancer. Mol Cancer. 2018;17(1):58.
27. Shyam Sunder S, Sharma UC, Pokharel S. Adverse effects of tyrosine kinase inhibitors in cancer therapy: pathophysiology, mechanisms and clinical management. Signal Transduct Target Therapy. 2023;8(1):262.

Figures

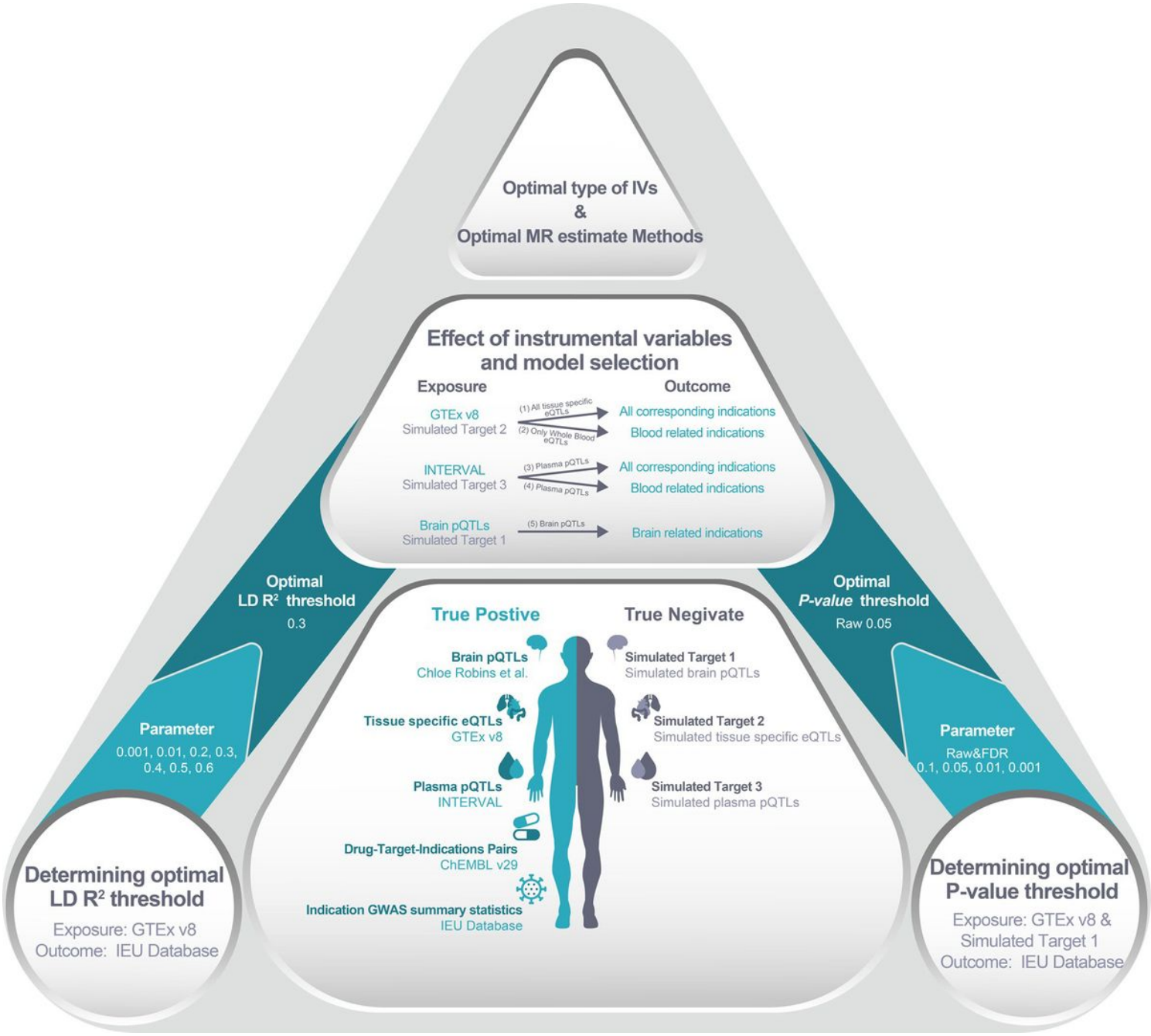

Figure 1

Study design.

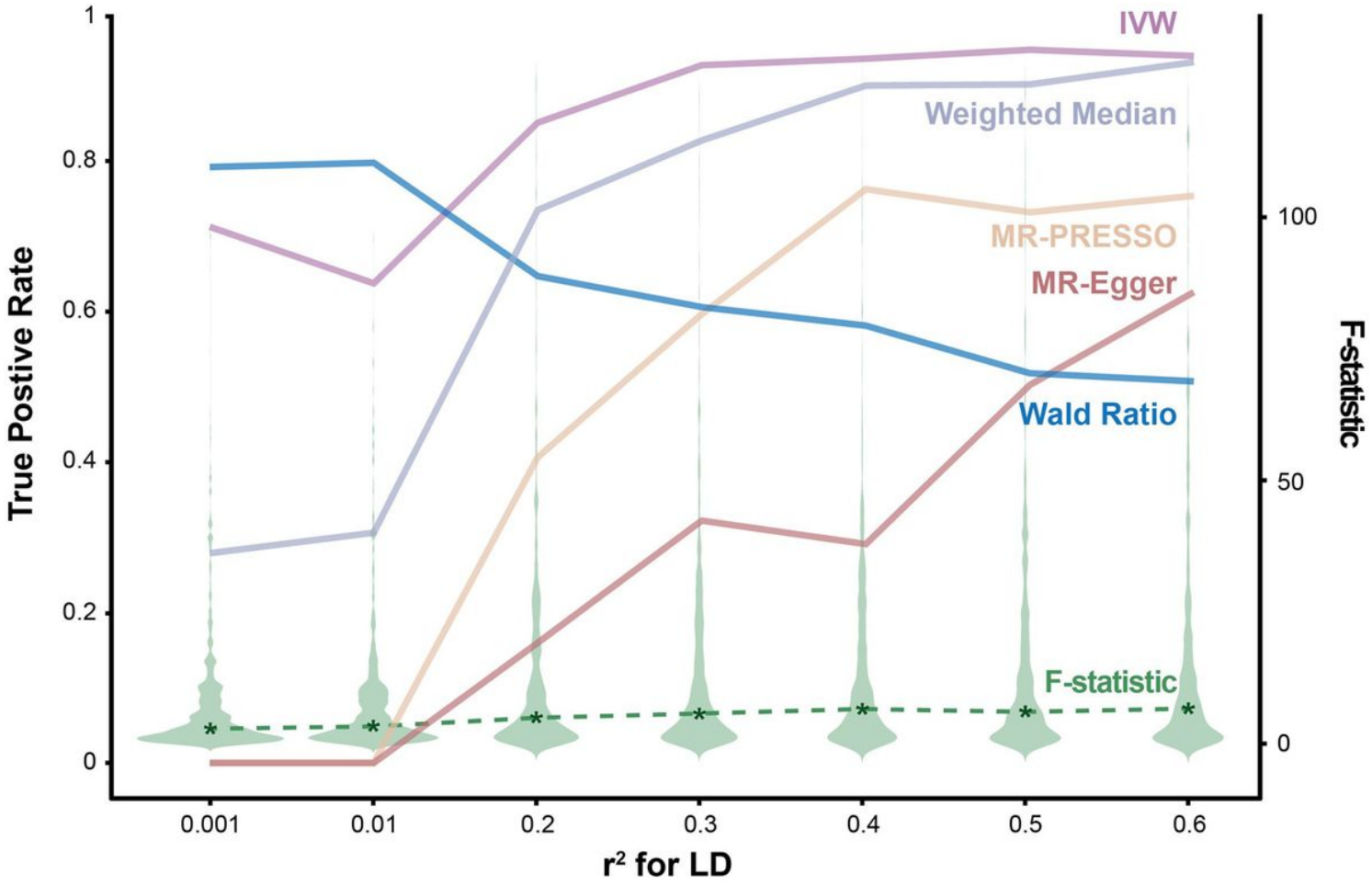

Figure 2

Evaluation metrics of LD r^2. The horizontal coordinate is the threshold of different r^2 values used by LD. The line chart refers to the left ordinate to show the true positive rate of different algorithms with the change in LD r^2, and the violin chart refers to the right Y-axis to show the change in the statistical efficacy of instrumental variables with the change in LD r^2.

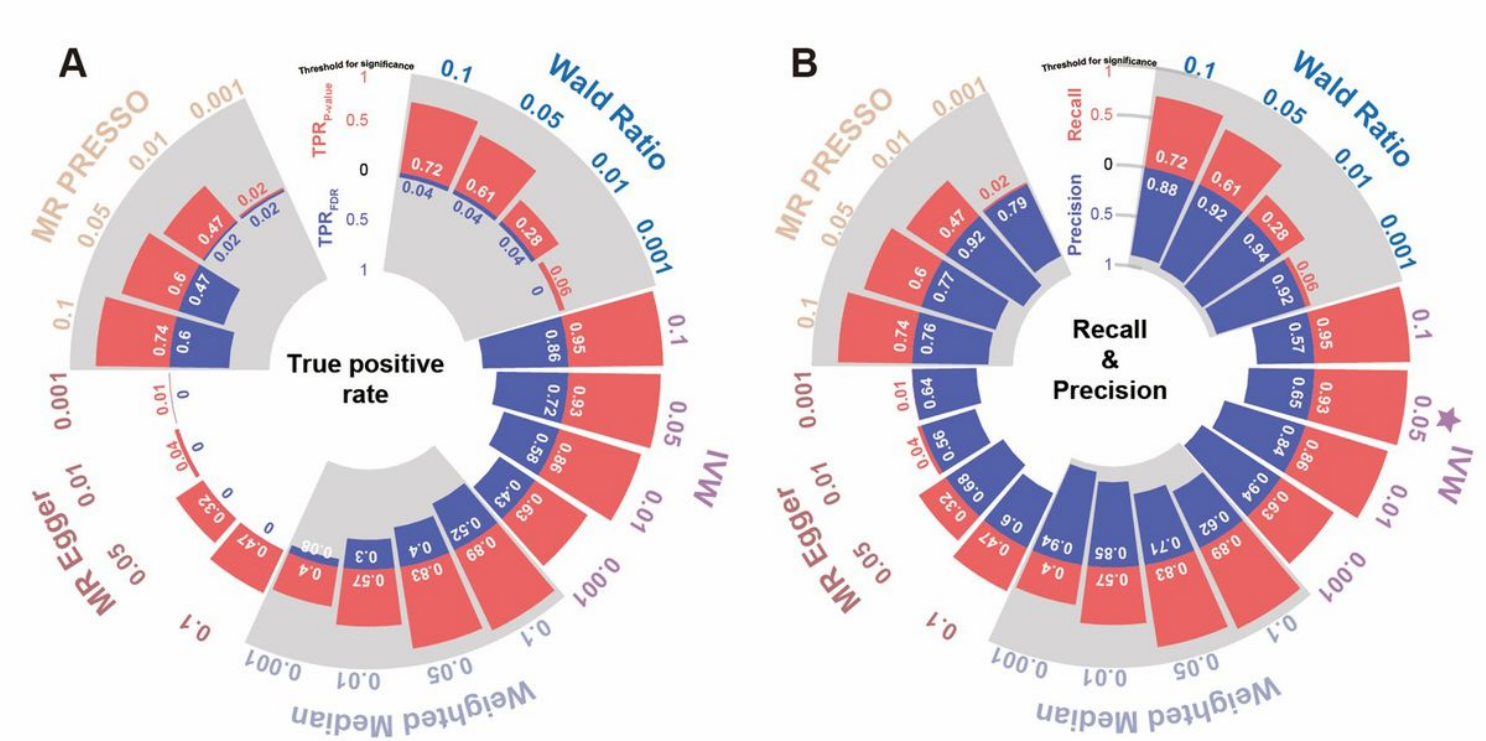

Figure 3

Evaluation metrics of the P value threshold for the results. Figure A shows the true positive rates of different models without FDR correction and after FDR correction under different P value thresholds. Figure B shows the recall and precision of different models without FDR correction under different P thresholds.

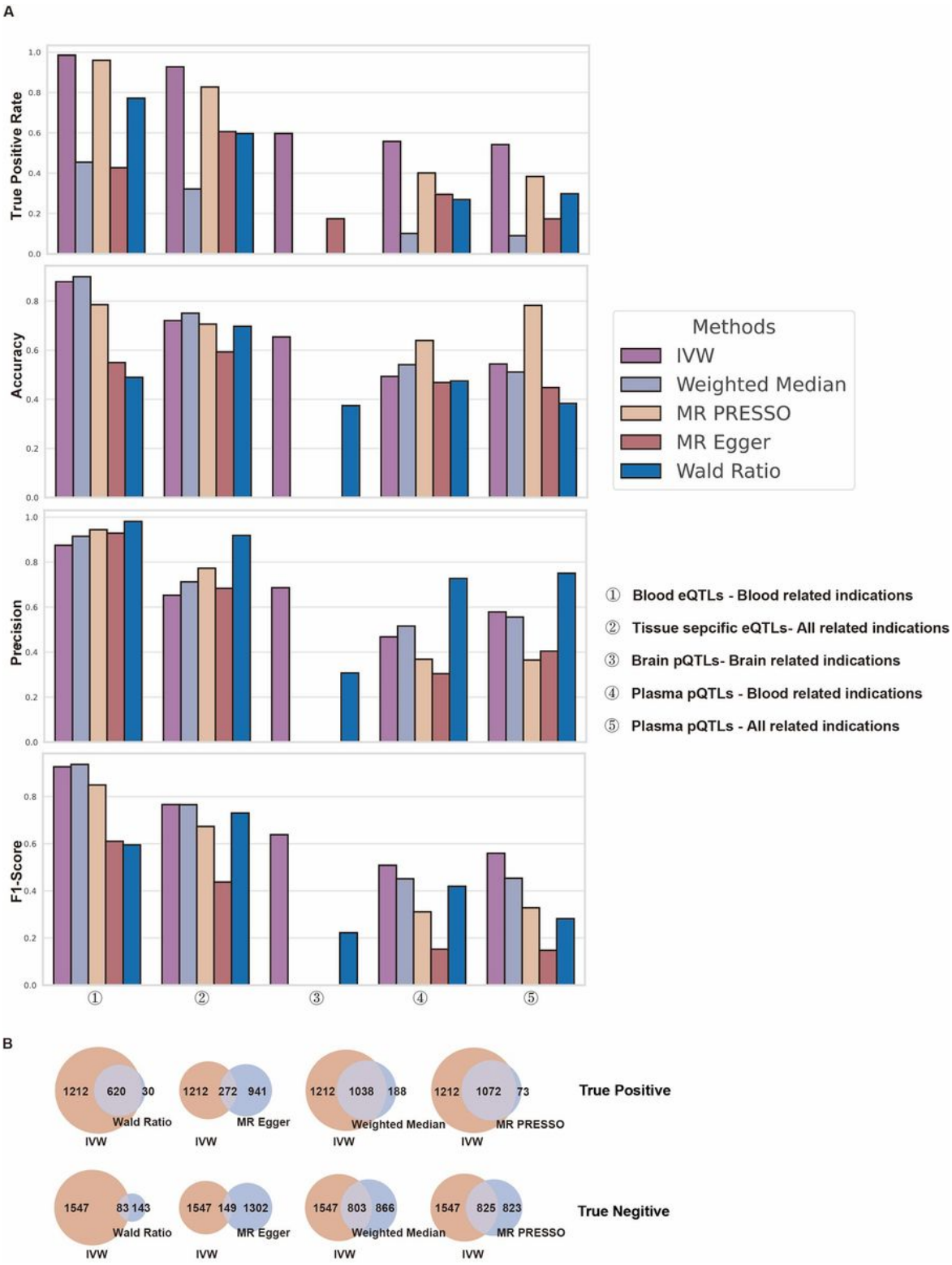

Figure 4

Evaluation metrics of instrumental variables and model selection on the effectiveness of drug MR. Figure A Five different datasets, true positive rate, accuracy rate, accuracy rate and F1-score changes in different algorithms. Figure B shows the number of overlaps of true positive and true negative results between IVW and other algorithms in the tissue-specific eqtl dataset.

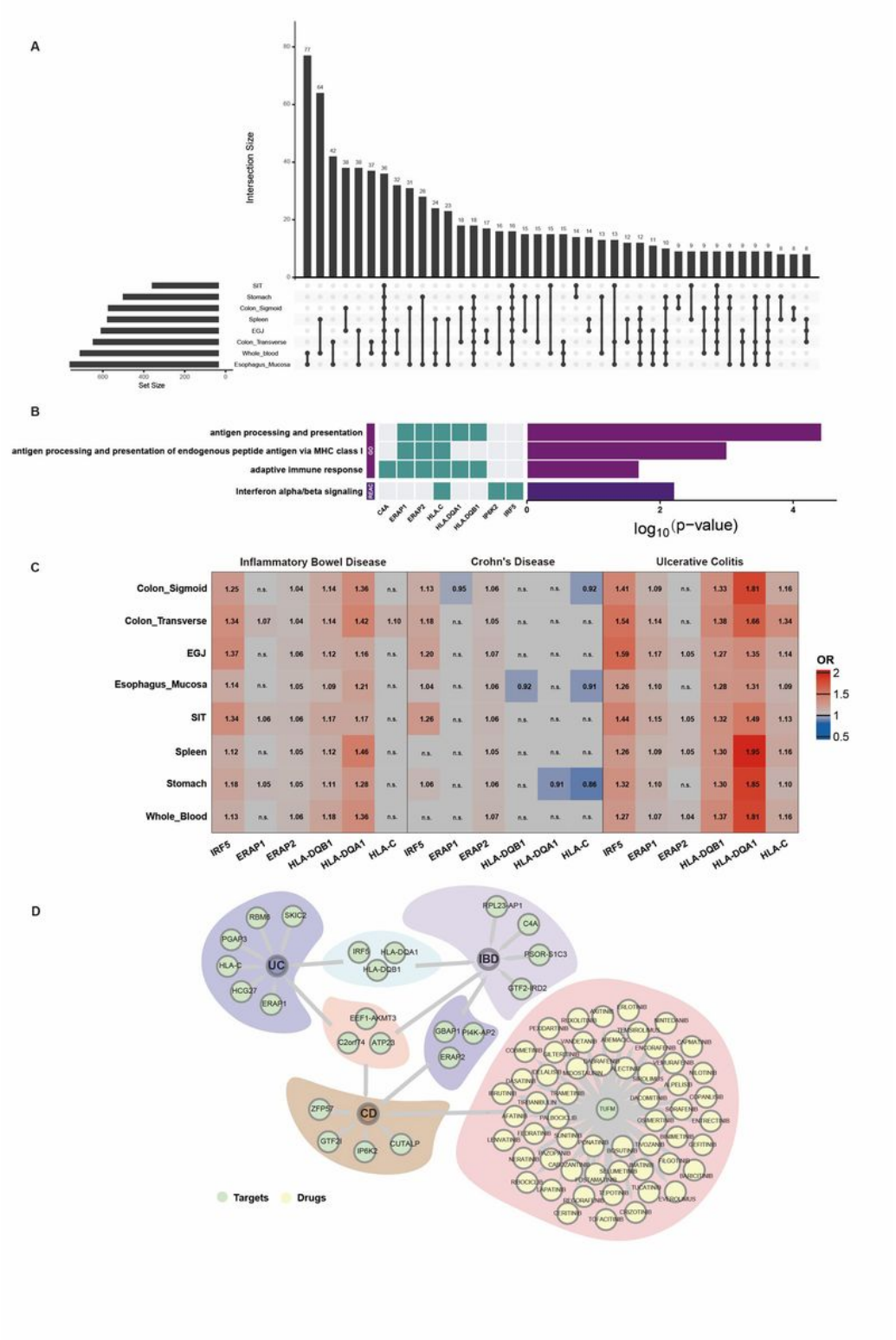

Figure 5

A. Upset plot. The left side of the figure shows the number of positive sample pairs in each tissue, the right side represents sample pairs shared among different tissues, and the top side represents the total number of sample pairs shared by that tissue. B. Pathway enrichment analysis results of genes in the pairs with positive results that occurred in all tissues. C. Heatmaps of the ORs of several new target genes. A red block indicates that inhibiting the gene has an effect on the disease, and a blue block

indicates that activating the gene has an effect on the disease. D. A network between novel genes and associated drugs. (SIT: small intestine terminal ileum, EGJ: Esophagus Gastroesophageal Junction)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ST11120.xlsx
- ST21120.xlsx
- ST31120.xlsx
- ST41120.xlsx
- ST5WholeBloodmergeMRres.xlsx
- ST6ColonSigmoidmergeMRres.xlsx
- ST7ColonTransversemergeMRres.xlsx
- ST8EsophagusGastroesophagealJunctionmergeMRres.xlsx
- ST9EsophagusMucosamergeMRres.xlsx
- ST10SmallIntestineTerminalIleummergeMRres.xlsx
- ST11SpleenmergeMRres.xlsx
- ST12StomachmergeMRres.xlsx
- ST13PostivateTargetwithdrug.csv
- SupplementaryFigure1.jpg