nature portfolio

Peer Review File

Multi-omics insight into the metabolic and cellular characteristics in the pathogenesis of hypothyroidism

Corresponding Author: Professor Yujuan Shan

Figures originally included in the author's rebuttal have been redacted from this file.

Attachments originally included by the reviewers as part of their assessment appear at the end of this file.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Shedding light on the role of immunity and inflammatory features of hypothyroidism is an interesting challenge. Genetic studies can provide insights to that matter. In this work, an effort to integrate multiple omics levels is made and some interesting results come out of this integration, including drug repurposing possibilities. However, to the reading of this manuscript, the reader could be first confused by the rather complicated selection pipeline. Moreover, statistical pipelines would benefit from changes to ensure robustness and coherence of the causal interpretation of the results. Also, the validity of mediation analyses is at stake as it stands. Here are recommendations to enhance the manuscript's transparency and clarity as well as ensuring a better robustness and result reproducibility.

Figures 2a, 3a and 4a each present similar if not identical pipelines. The formulation of the second block is confusing because the reader could understand that features are eliminated if the IVW p-value is higher than 5e-8. That is of course not the good interpretation. A clearer formulation would be appreciated to ensure coherence with the methods.

The large scale of recent genetic studies provides enough statistical power to avoid using Wald Ratios. This method is prompt to false positives since no control for horizontal pleiotropy can be provided. Despite this fact, the authors present single IV MR results with the same confidence as results from multiple IV MR techniques. The authors use a very stringent p-value threshold for IV selection. The validity of the first MR assumption is thus very well respected, but the number of IVs is insufficient to control potential bias caused by a violation of the two other assumptions. I'd suggest relaxing the imposed p-value threshold (e.g. 1e-6) for IV selection to increase the number of IVs. Systematically rejecting MRs with less than 3 IVs should conserve sufficient IV strength (F-statistic > 10) and ensure better control of MR assumptions 2 and 3 through the Weighted median estimator and Egger-intercept.

Moreover, many new findings (not already known in the literature) are made using a single IV : 1-oleoyl-GPC (18:1), 2-palmitoyl-GPC (16:0), 1-(1-enyl-palmitoyl)-2-arachidonoyl-GPE (p-16:0/20:4), 1-(1-enyl-palmitoyl)-2-oleoyl-GPE (p-16:0/18:1), CD25 on the surface of CD45RA+ CD4 not regulatory T cells, CD11c+ HLA DR++ monocyte %monocyte, lymphocyte absolute count, TNFSF11, TNFRSF11B. How would the authors defend the causal interpretation of these findings as opposed to the hypothesis of horizontal pleiotropy-driven estimated effect?

Also, minimal MAF for chosen IVs is 0.01. It is a little low given the size of the 1000 Genomes European subset (0.01 * 503 * 2 ≈ 10 occurrences of minor allele). (The size of this subset should also be specified in the methods.) The estimated r2 is thus not always precise enough given the 0.001 r2 threshold. I'd suggest using MAF>0.05 instead.

At line 69, please be precise in key point brought by reference 5. "…, the administration of levothyroxine does not yield to significant clinical improvements in elderly individuals with subclinical hypothyroidism."

At line 125, please precise the FDR calculation procedure used (e.g. Benjamini-Hochberg, Benjamini-Yekutieli) or add it to the methods. Also, please add an FDR adjusted p-value column in the supplementary data (S2, S7, S12, S17).

Line 125 and in the methods. Please be precise which sensitivity analysis were used for selection and at which threshold. From my understanding, significance of effect in weighted median (WM), Egger, RAPS and others were not mandatory and

were barely mentioned in the manuscript. How did those influence the authors' choices?

Line 129 and methods. No method is provided for the meta-analysis of discovery cohort MR analysis and replication cohort MR analysis. Why did the authors choose to do this meta-analysis instead of selecting only replicating results at nominal p-value?

Figure 2b does not represent discovery results and shows nonsignificant results. It would be more appropriate to use only retained causal features in the figures.

Why does it seem like there are missing results in Figure 3b in the replication and combined columns? Also, maybe a forest plot showing only retained causal features would be less confusing. Why a balloon plot here and a Forest plot in the previous figure for MR results?

Line 160. It is said that IVs chosen from cis-pQTL are better than trans-pQTLs (especially to ensure validity of first MR assumption) but is it what is done in the next section? The methods do not give a window size for cis-IV selection. If this comment concerned immune traits, many trans-QTLs could act on the exposure as other mechanisms than transcription are involved.

Line 165. Feature names including the “%” character can be confusing in a text.

It is not made very clear what the word interaction refers to at the section starting at line 183. No interaction terms are included in performed regressions.

The investigation of a potential causal link between TSH levels and hypothyroidism is highly subject to reverse causation and horizontal pleiotropy. Indeed, as the authors themselves highlight when referring to the literature, the disease often starts with lymphocyte infiltration in thyroiditis, which leads to a lower production of T4 hormone. Lower T4 levels result in a higher production of TSH. Control for reverse causation should thus be present. However, the authors did not provide sensitivity analysis for horizontal pleiotropy or even reverse causation for this part of the study. Also, the authors should specify if TSH GWAS only includes individuals with normal TSH levels and no thyroid disfunction background to avoid bias towards an association with hypothyroidism.

The mediation analysis lacks robustness and is poorly described in the methods. The method used is unclearly described and does not seem to make the difference between vertical and horizontal pleiotropy. No reference is present for Two-sample MR. The reference for the usage of the delta method itself refers to another paper which does not use this method in a mediation analysis context but rather for Wald Ratio estimator.

At lines 292-294, a reference is given to support the finding of TNFSF11 as a protective factor. The referenced study also used a single IV (undoubtedly the strongest QTL found also in the outcome data) and performed the analysis using the same GWAS of outcome as for the Discovery cohort (FinnGen) and the same pQTL summary statistics as exposure. The obtained result was thus predictably the same as in this other study and cannot be considered new.

GWAS IDs are not provided in the section starting at line 349.

Line 385. The reference 46 does not seem to be the original reference to the proposed r2 approximation. Please provide the original reference for the proposed method.

Lines 393-397. How to know which from fixed or random effect was kept for each MR in the supplementary data?

Line 409. How are cases of Heterogeneity defined? Which test? Which threshold?

Lines 414-416. The MR assumption 2 is not well formulated. It should be: “Genetic variants are not associated with a confounding factor between exposure and outcome”.

Cell type-specific enrichment analysis could be better explained. What are TAG sets? Is it needed to provide a list of genes/features as an input? What does this analysis bring to the manuscript that was not already known, namely the implication of lymphocytes in the disease?

Lines 375-376. What are the “unknown metabolites”?

Lines 117-118. Did the authors mean that F statistic of exposure ~ genetic variant should exceed 10? It does not make sense otherwise. Also, to which F-statistic does the F column refer to in the supplementary tables?

According to the study's results, TNFSF11 acts as a protective factor for hypothyroidism. However, the proposed drugs (lines 235-236) act as inhibitors of the RANK/RANKL system. This appears to be a contradiction.

Reviewer #2

(Remarks to the Author)
This study aimed to assess the (causal) relationship of immunological and inflammatory proteins with hypothyroidism.

Authors analysed several data and used a range of well-known statistical genetic methods. To me, the subject of this study is relevant and important. Also, the methods used are well-regarded, and overall, the analysis appeared to have been rigorously conducted. However, it occurs to me that the study has not been well presented coupled with several issues around grammatical errors. This study may benefit from professional language editing. Here are some specific comments for authors to consider in improving the quality and presentation of their study.

Abstract
1. Lines 46 – 48 which were meant to set the tone of the study were not well constructed. Authors aimed to provide a rationale for their study, but this did not come across clearly. Please, rewrite the sentence.

Introduction
1. I will recommend a revision of this introduction section, with a particular focus on clarity of expression, and grammatical accuracy. Authors need to rewrite a substantial part of this introduction to meet publication standards. There may be a need for professional English language editing to improve the presentation of this work. For example, 'Given that the heterogeneity of the elderly population, it is imperative to formulate new preventive strategies against the disease' (lines 69 – 71) is not clear. The same comment applies to lines 72 – 75, line 76 ('investigated'), line 86 ('lower susceptibility'), etc. In line 87, 'significant' should be used in relation to statistical assessment.

2. Authors made an effort to present their study, however, they have not provided sufficient justification or a convincing rationale for it. This issue may be related to the clarity of writing, which the authors can work on improving. In my opinion, addressing this matter is crucial for the paper to progress.

3. Lines 89 – 90, several studies have been conducted on this topic. What is the summary of the findings of those studies and how does the current study contribute or add something new? In other words, what are the limitations or gaps in knowledge that your study seeks to fill?

Results
1. It is not clear why only very few instruments were used for the Mendelian randomisation analysis, for example, for the IVW model (Supplementary Table S9, for example). How reliable are the results based on these limited instruments? If there are not enough genome-wide significant SNPs, could authors have relaxed the cut-off point to say, genome-wide suggestive level?
2. Legend on Figure 2C is somewhat confusing, as it came across as if PH4 = 0.8 represents no support. Please, consider a correction here. This comment also applies to Figure 3B.

3. Lines 180 – 181, do authors mean the TNFSF11 level was causally protective for hypothyroidism?

4. Line 209, .'.. still remain to be fully explored.' Are there studies that have looked at them? if yes, cite them appropriately, and show what the current study is adding.
Discussion
The discussion section is also not well presented. Authors need to do a bit of more work here to improve this section. Please, check how previous studies discuss their findings. Some specific points here are listed below.
1. Line 244, '…this work is among the first …' is rather ambiguous. If it is the first, say so, If there are previous studies, authors need to acknowledge, cite them appropriately and state how their work advances those.
2. Line 248 'GPL'. Authors used this term severally without first defining it. Please, define all terms, abbreviations, etc, at first use.
3. Line 250 'This result was confirmed through …' Is it confirmed by the study, or your findings confirmed the study, or your findings align/agree with ...?
4. Grammar/punctuation errors, line 262 ('role in both These')
5. I was expecting information on how the identified 'targets' could be repurposed (noted by authors in lines 232 – 241) in the discussion section.

Methods
- Lines 351 – 354, Authors chose the GWAS data with a lower sample size as the discovery set, I have no problem with this, but any justification for such a decision? Also, is it possible to put all the data descriptions under one heading, say 'GWAS data source' rather than spreading them out under different sub-headings?
- Lines 376 – 377, 'we additionally excluded those unknown metabolites to obtain more definitive association results.' Is not clear.
- In line 399, 'MR requirements' should be 'MR assumption'.

Reviewer #3

(Remarks to the Author)
Zheng and colleagues apply mendelian randomization and colocalization analysis to identify ten putative causal traits of hypothyroidism by leveraging multi-omics data including inflammatory proteins, plasma metabolites and immune cell traits. Additionally, the following-up mediation analysis reveal the potential mediating role of thyroid-stimulating hormone and the cell type-specific enrichment analysis demonstrate the important role of lymphocytes in the pathogenesis of hypothyroidism. Overall, this manuscript is well-organized, offering a clear and logical structure that enhances understanding. However, I also found multiple areas for improvement in the manuscript, including (a) additional results and interpretation of the

sensitivity in the results presented, (b) additional clarity in the methods presented, (3) additional figures for the colocalization analysis and (4) lack of analysis in the non-European population.

I outline my detailed comments below.

1. Abstract: I suggest that the authors could incorporate some information regarding the identified putative causal traits within the text of the Abstract since these putative causal traits are the most important findings in this manuscript.

2. Results - Mendelian Randomization: I suggest that the authors could incorporate some results and interpretations of sensitivity analysis after each Mendelian Randomization, for example, Cochran's Q statistic, MR-Egger regression and MR-PRESSO, which are the methods that authors pointed out at the Methods section and supplementary data.

3. Results – P value threshold: in the section of 'Associations between plasma metabolites and hypothyroidism ', the authors used FDR as the multiple testing correction, while in the section of 'Association between immune cell traits with hypothyroidism', the authors used Bonferroni method as the multiple testing correction. Could the authors provide with the reasons using two different multiple testing correction methods?

4. Methods/Results – Colocalization analysis: I suggest that the authors could incorporate the statistical fine-mapping (SuSiE) into the colocalization analysis, which can relax the single causal variants assumption. [Wallace C. A more accurate method for colocalisation analysis allowing for multiple causal variants. PLoS Genet. 2021 Sep 29;17(9):e1009440. doi: 10.1371/journal.pgen.1009440. PMID: 34587156; PMCID: PMC8504726.]

5. Methods – Study design: I am curious about is there any specific reasons that the authors used FinnGen R9 study as the discovery study, while use another data as the replication study. Based on the sample size the authors showed at the Method section, the second data has more cases and controls compared to the FinnGen R9 study. Additionally, I also suggest that the authors may conduct the meta-analysis by combining these two studies to enhance the power for the following analyses.

6. Methods – Colocalization analysis: I suggest that the authors could clarify how they set up the window size/region to select the variants for each colocalization analysis.

7. Methods/data sources: The data sources that the authors used for Mendelian Randomization and Colocalization analysis were from European populations, which might limit the reliability when extrapolating the findings to non-European populations. I recommend that the authors could use available non-European data sources.

8. Methods/data sources: The data sources that the authors used for Mendelian Randomization and Colocalization analysis were from European populations, which might limit the reliability when extrapolating the findings to non-European populations. I recommend that the authors could use available non-European data sources.

9. Figure2c/3c – Colocalization analysis: I suggest that the authors could generate the colocalization plot for the colocalization analysis, which is helpful for the interpretation of the colocalization result. Here is R function to create the coloc plot: locuscompare(). https://hanruizhang.github.io/GWAS-eQTL-Colocalization/

Author Rebuttal letter:

Dear reviewers,

RE COMMSBIO-24-0721

Multi-omics insight into the molecular and cellular characteristics in the pathogenesis of hypothyroidism

We would like to thank you for your helpful comments and suggestions for improving our manuscript. We have uploaded a clean version of the revised manuscript and a version with changes in red color. Our point-by-point responses to your comments are presented below in blue color.

Reviewer #1:

Shedding light on the role of immunity and inflammatory features of hypothyroidism is an interesting challenge. Genetic studies can provide insights to that matter. In this work, an effort to integrate multiple omics levels is made and some interesting results come out of this integration, including drug repurposing possibilities. However, to the reading of this manuscript, the reader could be first confused by the rather complicated selection pipeline. Moreover, statistical pipelines would benefit from changes to ensure robustness and

coherence of the causal interpretation of the results. Also, the validity of mediation analyses is at stake as it stands. Here are recommendations to enhance the manuscriptâs transparency and clarity as well as ensuring a better robustness and result reproducibility.

Response: We thank you for your helpful comments and interest in our research topics. In our revision version, we took your suggestions and simplified the selection pipeline to increase readability. We also modified the statistics pipelines including the sensitivity analysis to enhance the robustness of the results. In addition, we re-evaluated the validity of the mediation analysis. Please see our responses to your specific comments below:

Figures 2a, 3a and 4a each present similar if not identical pipelines. The formulation of the second block is confusing because the reader could understand that features are eliminated if the IVW p-value is higher than 5e-8. That is of course not the good interpretation. A clearer formulation would be appreciated to ensure coherence with the methods.

Response: Sorry for the potential confusion caused. We have now modified the selection pipeline and formulated a uniform selection pipeline for all exposures, as shown in new Figure 1. We also changed the description of the instrumental variable selection to reduce confusion during reading, such as replacing âP-value > 5 Ã10 -8â with âExtracted SNPs that associated with exposure (P < 5Ã10-6 and minor allele frequency > 0.05)â in new Figure 1.

The large scale of recent genetic studies provides enough statistical power to avoid using Wald Ratios. This method is prompt to false positives since no control for horizontal pleiotropy can be provided. Despite this fact, the authors present single IV MR results with the same confidence as results from multiple IV MR techniques. The authors use a very stringent p-value threshold for IV selection. The validity of the first MR assumption is thus very well respected, but the number of IVs is insufficient to control potential bias caused by a violation of the two other assumptions. Iâd suggest relaxing the imposed p-value threshold (e.g. 1e-6) for IV selection to increase the number of IVs. Systematically rejecting MRs with less than 3 IVs should conserve sufficient IV strength (F-statistic > 10) and ensure better control of MR assumptions 2 and 3 through the Weighted median estimator and Egger-intercept.

Response: We agree with your concern about pleiotropy causing false positives in our results. As you suggested, we adjusted the p-value threshold of instrumental variable selection from 5e-8 to 5e-6 and excluded those exposures with less than three SNPs.

Remarkably, although we relaxed the p-value threshold, for inflammatory proteins, only a very small number of inflammatory proteins had at least three SNPs. Considering that we did not conclude novel results in the original version, we removed GWAS data for inflammatory proteins in the revised version and only focused on the causal effects of metabolites and immune cells on hypothyroidism.

At this new p-value threshold (5e-6), nearly 96% of exposures, including metabolites and immune cells traits, had at least three SNPs for subsequent analysis. We believe that the sensitivity analysis performed at the new p-value threshold will allow for better control of the second and third hypotheses of MR.

Lines 157-161: âAt the LD clumping criteria of pairwise LD R 2 < 0.001 within a 10,000 kb window and the significance threshold of P-value < 5Ã10 -6, nearly 96% of exposures (1,150 exposures for plasma metabolites and 683 exposures for immune cells traits) had at least three SNPs for subsequent analysis (Supplementary Data 2 and 3)â (Results section)

Lines 580-582: âWe also systematically removed exposures that had fewer than three IVs, thereby ensuring the conduction of subsequent pleiotropy testâ (Methods section)

Moreover, many new findings (not already known in the literature) are made using a single IV : 1-oleoyl-GPC (18:1), 2-palmitoyl-GPC (16:0), 1-(1-enyl-palmitoyl)-2-arachidonoyl-GPE (p-16:0/20:4), 1-(1-enyl-palmitoyl)-2-oleoyl-GPE (p-16:0/18:1), CD25 on the surface of CD45RA+ CD4 not regulatory T cells, CD11c+ HLA DR++ monocyte %monocyte,

lymphocyte absolute count, TNFSF11, TNFRSF11B. How would the authors defend the causal interpretation of these findings as opposed to the hypothesis of horizontal pleiotropy-driven estimated effect?

Response: Compared to the original version, we have made more efforts to control pleiotropy. First, for the selection of IVs, we further introduced RadialMR method (Bowden J, et al, Int J Epidemiol. 2018;47(4):1264-1278), which can identify outlier pleiotropic SNPs. Before MR analysis, we used this method to exclude outlier pleiotropic SNPs in IVs to reduce the risk of pleiotropy.

Lines 582-585: âMoreover, we further introduced RadialMR method, which can identify outlier pleiotropic SNPs 35. Before MR analysis, we used this method to exclude outlier pleiotropic SNPs in IVs to reduce the risk of pleiotropyâ (Methods section)

Second, due to the adjusted p-value threshold, there are sufficient IVs (nâ¥3) used for the pleiotropy test, including Egger intercept test and MR-PRESSO global test. We believe these efforts will reduce the threat of pleiotropy to our findings.

Also, minimal MAF for chosen IVs is 0.01. It is a little low given the size of the 1000 Genomes European subset (0.01 * 503 * 2 â 10 occurrences of minor allele). (The size of this subset should also be specified in the methods.) The estimated r2 is thus not always precise enough given the 0.001 r2 threshold. Iâd suggest using MAF>0.05 instead.

Response: Thank you for the suggestion. We have adjusted the MAF for MR analysis to at least 0.05. We have also added a description of the European sample size of 1000 genomes in the methods section.

Lines 572-576: âFor each exposure, we initially filtered SNPs with a threshold of P-value < 5Ã10-6 and minor allele frequency (MAF) > 0.05. Subsequently, we performed clumping by the PLINK software using a cut-off of R2 < 0.001 and window size of 10,000 kb, utilizing the 503 European samples data from the 1000 Genomes project as our reference panel 34â (Method section)

At line 69, please be precise in key point brought by reference 5. ââ¦, the administration of levothyroxine does not yield to significant clinical improvements in elderly individuals with subclinical hypothyroidism.â

Response: Thank you for the suggestion. According to the suggestion provided by reviewer 2, we have extensively revised the introduction section in order to enhance the clarity of expression. The sentence you pointed out has been deleted in the revised version.

At line 125, please precise the FDR calculation procedure used (e.g. Benjamini-Hochberg, Benjamini-Yekutieli) or add it to the methods. Also, please add an FDR adjusted p-value column in the supplementary data (S2, S7, S12, S17).

Response: In the previous version, we adjusted for multiple tests in two methods, including Bonferroni method and FDR under BenjaminiâHochberg method, which could potentially lead to confusion. In the revised version, we consistently used the Bonferroni method for multiple test correction across all analyses. Therefore, we did not add extra FDR adjust p-value column in the supplementary data.
Lines 170-177: âAfter adjustment for multiple testing by Bonferroni method (P-value < 4.35Ã10-5), we identified that 17 plasma metabolites were significantly associated with hypothyroidism risk using inverse-variance weighted method (Fig. 2a and Supplementary Data 4)â (Results section)

Lines 226-230: âAfter adjustment for multiple testing by Bonferroni method (P-value < 7.32Ã10-5), we identified a total of 26 immune cell traits that were significantly associated with hypothyroidism using inverse-variance weighted method (Fig. 3a and Supplementary Data 7)â (Results section)

Lines 608-612: âAdjustment for multiple testing was conducted to avoid the potential

influence of false positive findings. Bonferroni method was used across all MR analysesâ (Methods section)

Line 125 and in the methods. Please be precise which sensitivity analysis were used for selection and at which threshold. From my understanding, significance of effect in weighted median (WM), Egger, RAPS and others were not mandatory and were barely mentioned in the manuscript. How did those influence the authors' choices?

Response: Thank you for the suggestion. In the revision version, we provide a detailed explanation of how analysis result affects our judgment. A significant association was considered confident if it met a series of stringent criteria: (1) the significance of association passed multiple tests using IVW method as well as the directional consistency of the effect using other extra four MR approaches, (2) there was no significant heterogeneity among IVs (P for Cochranâs Q > 0.05), (3) there was no evidence of horizontal pleiotropy (Egger P for intercept > 0.05 and MR-PRESSO global test P > 0.05) and (4) no reverse causal MR effect (P for MR Steiger test > 0.05).

Lines 180-186: âAmong them, the association between kynurenine and hypothyroidism was not passed the sensitivity analysis due to the inconsistent effect direction of the additional method (MR Egger) (Supplementary Data 4). Other 16 significant associations were not biased by horizontal pleiotropy (Egger P for intercept > 0.05 and MR-PRESSO global test P > 0.05), heterogeneity among IVs (P for Cochranâs Q > 0.05) or reverse causation (P for MR Steiger test > 0.05) (Supplementary Data 4)â (Results section)

Lines 233-236: âAmong them, two associations of TCRgd T cell to lymphocyte ratio and CD28- CD127- CD25++ CD8 + T cell to T cell ratio with hypothyroidism were not passed the sensitivity analysis due to the inconsistent effect direction of the additional method (MR Egger) (Supplementary Data 7)â (Results section)

Lines 347-351: âThree associations among them were not passed the sensitivity analysis due to the inconsistent effect direction of the additional method (MR Egger), which were 1-(1-enyl-palmitoyl)-2-oleoyl-GPE (p-16:0/18:1)-chronic kidney disease, CD27 on IgD-CD38+ B cell-heart failure and CD3 on CD39+ CD4+ T cell-diabetic nephropathy pairs (Supplementary Data 13)â (Results section)

Lines 645-651: âIn our study, a significant association was considered confident if it met a series of stringent criteria: (1) the significance of association passed multiple tests using IVW method as well as the directional consistency of the effect using other extra four MR approaches, (2) there was no significant heterogeneity among IVs (P for Cochranâs Q > 0.05), (3) there was no evidence of horizontal pleiotropy (Egger P for intercept > 0.05 and MR-PRESSO global test P > 0.05) and (4) no reverse causal MR effect (P for MR Steiger test > 0.05)â (Methods section)

Line 129 and methods. No method is provided for the meta-analysis of discovery cohort MR analysis and replication cohort MR analysis. Why did the authors choose to do this meta-analysis instead of selecting only replicating results at nominal p-value?

Response: In our study, the associations for each exposure on two outcome data were meta-analyzed using the random effects model.

Lines 197-200: âIn the meta-analysis of discovery and replication datasets using random effect model, five plasma metabolites displayed nominal significant associations with hypothyroidism (P-value < 0.05) (Fig. 2b and Supplementary Data 6)â (Results section)

Lines 238-241: âIn the meta-analysis of discovery and replication datasets using random effect model, 15 immune cell traits displayed nominal significant associations with hypothyroidism (Fig. 3b and Supplementary Data 9)â (Results section)

Lines 616-618: âMeta-analyses of MR result in both the discovery and replication cohorts were calculated using the random effect model (R package meta, version 7.0-0)â (Methods section)

In fact, integrating MR results using meta-analysis from multiple independent outcome

cohorts is very common in current MR studies (Zou X, et al, CNS Neurosci Ther, 2023, 29(10):3043-3052; Chen J, et al, Diabetes Care. 2023, 46(4):828-835; Yuan S, et al, Cell Rep Med. 2023, 4(9):101174), which will ensure sufficient statistical power and robustness of the study results.

Figure 2b does not represent discovery results and shows nonsignificant results. It would be more appropriate to use only retained causal features in the figures.

Response: Thank you for the suggestion. We have now only retained causal features in the new Figure 2 and 3.

Why does it seem like there are missing results in Figure 3b in the replication and combined columns? Also, maybe a forest plot showing only retained causal features would be less confusing. Why a balloon plot here and a Forest plot in the previous figure for MR results?

Response: Sorry for the potential confusion caused. In the previous version, the size of the bubble in Figure 3b corresponds to the P-value. The presence of missing result indicates that the P-value lacks statistical significance. We have now consistently used forest plot to present our results to reduce confusion in the new Figure 2 and 3.

In the new Figure 2 and 3, Figure 2a and Figure 3a demonstrate significant causal associations between plasma metabolites and immune cell traits and hypothyroidism at the discovery stage, respectively. Figure 2b and Figure 3b showed significant causal associations of plasma metabolites and immune cell traits with hypothyroidism after MR meta-analysis, respectively.

Line 160. It is said that IVs chosen from cis-pQTL are better than trans-pQTLs (especially to ensure validity of first MR assumption) but is it what is done in the next section? The methods do not give a window size for cis-IV selection. If this comment concerned immune traits, many trans-QTLs could act on the exposure as other mechanisms than transcription are involved.

Response: Sorry for the potential confusion caused. cis-pQTLs were defined as pQTLs located within 1 Mb around a transcription start site of a protein-coding gene as described in the previous literature (Yoshiji S, et al, Nat Metab. 2023, 5(2):248-264). In our study, some immune traits belong to protein, such as CD3 on CD4 regulatory T cell, and no cis-pQTL in the related IVs. Although trans-pQTLs is a greater threat to the horizontal pleiotropy of MR results than cis-pQTLs, the pleiotropy test performed in our study can avoid this concern to some extent. Therefore, in our revised version, we still retained these significant associations for subsequent analyses, rather than excluding them as we did in the original version.

Line 165. Feature names including the â%â character can be confusing in a text.

Response: Sorry for the potential confusion caused. âA%Bâ means the âA to B ratioâ. We have rewritten the exposure names containing â%â character to reduce potential confusion in the revised version.

It is not made very clear what the word interaction refers to at the section starting at line 183. No interaction terms are included in performed regressions.

Response: Sorry for the potential confusion caused. We found that the word "interaction" in the original version was indeed misleading. In the revised version, we removed the content associated with it to avoid confusion.

The investigation of a potential causal link between TSH levels and hypothyroidism is highly subject to reverse causation and horizontal pleiotropy. Indeed, as the authors themselves highlight when referring to the literature, the disease often starts with lymphocyte infiltration in thyroiditis, which leads to a lower production of T4 hormone. Lower T4 levels result in a higher production of TSH. Control for reverse causation should thus be present. However, the authors did not provide sensitivity analysis for horizontal pleiotropy or even reverse causation for this part of the study. Also, the authors should

specify if TSH GWAS only includes individuals with normal TSH levels and no thyroid disfunction background to avoid bias towards an association with hypothyroidism.

Response: Thank you for the suggestion. In the previous version of the causal analysis of TSH level on hypothyroidism, we did omit to assess the effect of horizontal pleiotropy and reverse causality on MR result. We did a sensitivity analysis again. The result was consistent with your point that MR Steiger test showed evidence of reverse causation and there was significant heterogeneity among the IVs. These results did not pass sensitivity analysis. Therefore, we removed TSH from the revision version, along with subsequent mediation analysis. Thank you for pointing out this serious mistake again.

The mediation analysis lacks robustness and is poorly described in the methods. The method used is unclearly described and does not seem to make the difference between vertical and horizontal pleiotropy. No reference is present for Two-sample MR. The reference for the usage of the delta method itself refers to another paper which does not use this method in a mediation analysis context but rather for Wald Ratio estimator.

Response: Thank you for pointing out this serious mistake. In the previous version, a two-step network MR analysis was performed to explore the potential mediation effect (Carter A, et al, Int J Epidemiol, 2021, 36(5):465-478).

Two MR estimates are calculated i) the causal effect of the exposure on the mediator (Î²1) and ii) the causal effect of the mediator on the outcome (Î²2) using univariate MR. These two estimates can then be multiplied together to estimate the indirect effect. The mediation proportion of each mediator in the associations between exposure and outcome was calculated as the product of Î²1 and Î²2 divided by the total effect of education on the outcome. The 95% CIs of the mediation proportions were calculated using the delta method (MacKinnon DP, et al, Psychol Methods, 2002, 7(1):83-104).

Due to the previously mentioned effect of reverse causality and heterogeneity on MR results, we removed TSH from our revision version, along with subsequent mediation analysis. Therefore, we did not add the modified statements about mediation analysis to the methods section in the revised version.

At lines 292-294, a reference is given to support the finding of TNFSF11 as a protective factor. The referenced study also used a single IV (undoubtedly the strongest QTL found also in the outcome data) and performed the analysis using the same GWAS of outcome as for the Discovery cohort (FinnGen) and the same pQTL summary statistics as exposure. The obtained result was thus predictably the same as in this other study and cannot be considered new.

Response: We agree with you that TNFSF11 is a protective factor for hypothyroidism cannot be considered a new conclusion. We also mentioned in the previous questions that there were few IVs of inflammatory proteins and we did not find any novel causal inflammatory proteins for hypothyroidism, so we removed the relevant content of inflammatory proteins in the revised version.

GWAS IDs are not provided in the section starting at line 349.

Response: Sorry for the potential confusion caused. The GWAS ID is applicable to GWAS data downloaded from the âopen GWAS projectâ, such as hypothyroidism complications in our study. However, the GWAS data for hypothyroidism in our study were derived from the websites of FinnGen R9 study and UK biobank, so there was no corresponding GWAS ID.

Line 385. The reference 46 does not seem to be the original reference to the proposed r2 approximation. Please provide the original reference for the proposed method.

Response: We have revised the reference for r2 approximation to the original reference in the revision version (Park JH, et al, Nat Genet, 2010;42(7):570-575).

Lines 591-593: âR2 is approximately calculated using the formula R2=2 Ã EAF Ã (1âEAF) Ã Î²2, where EAF represents the effect allele frequency, and Î² denotes the estimated

genetic effect on exposure 37â (Methods section)

Lines 831-833: â37. Park, J.-H. et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat Genet 42, 570â575 (2010)â (References section)

Lines 393-397. How to know which from fixed or random effect was kept for each MR in the supplementary data?

Response: Sorry for the potential confusion caused. We have specified specific analytical methods in the "Method" column of supplementary data.
Line 409. How are cases of Heterogeneity defined? Which test? Which threshold?

Response: Sorry for the potential confusion caused. MR estimates for each outcome from different sources were combined by the random-effects meta-analysis method. The I2 statistic was calculated to assess the heterogeneity of each outcome from different data sources, and the I2 values <25%, 25â75%, and >75% were considered to indicate low, moderate, and high heterogeneity, respectively as described in the previous literature (Chen J, et al, Diabetes Care, 2023, 46(4):828-835). We have added these in the revised version.

Lines 619-622: âThe I2 statistic was calculated to assess the heterogeneity of each outcome from different data sources, and the I2 values <25%, 25â75%, and >75% were considered to indicate low, moderate, and high heterogeneity, respectively, as described in the previous literature 40â (Methods section)

Lines 414-416. The MR assumption 2 is not well formulated. It should be: âGenetic variants are not associated with a confounding factor between exposure and outcomeâ.

Response: Thank you for your advice on grammar and organization. We have modified this sentence in the revised version.

Lines 627-628: âSecond, genetic variants are not associated with a confounding factor between exposure and outcomeâ (Methods section)

Cell type-specific enrichment analysis could be better explained. What are TAG sets? Is it needed to provide a list of genes/features as an input? What does this analysis bring to the manuscript that was not already known, namely the implication of lymphocytes in the disease?

Response: Sorry for the potential confusion caused. We investigated the cell enrichment of hypothyroidism in thyroid tissue in the Cell type-Specific Enrichment Analysis DataBase (CSEA-DB, available at https:// bioinfo.uth.edu/CSEADB/) (Dai Y, et al, Nucleic Acids Res, 2021, 49(D1):D862-D870). CSEA-DB collected GWAS aggregate statistics from three main sets: the Multi-trait Set (MTC) panel; The British biological bank (UKBB) group; Extended Trait Collection (ETC). After quality control, trait-associated gene (TAG) sets for all collected GWAS studies were calculated by Multi-marker Analysis of GenoMic Annotation method. This does not require additional gene input by the user. The detailed calculation method could refer to the original literature (Dai Y, et al, Nucleic Acids Res, 2021, 49(D1):D862-D870). In the methods section of the revised version, we added further explanations on this part.

Lines 447-455: âThe Cell type-Specific Enrichment Analysis DataBase (CSEA-DB, https://bioinfo.uth.edu/CSEADB/) 49, was used to investigate the enrichment of 12 general cell types in thyroid tissue of hypothyroidism patients. CSEA-DB collected GWAS aggregate statistics from three main sets: the Multi-trait Set (MTC) panel; The British biological bank (UKBB) group; Extended Trait Collection (ETC). After quality control, trait-associated gene (TAG) sets for all collected GWAS studies were calculated by Multi-marker Analysis of GenoMic Annotation method. We used three hypothyroidism datasets for analysis. The trait IDs were B1020 (128 cases), B2212 (17,574 cases) and B322 (118 cases)â (Methods section)

In our MR analysis, the GWAS of immune cell was conducted based on peripheral blood samples. However, there is currently a lack of available GWAS data for immune cells at the tissue level (i.e. thyroid tissue). Although it does not establish precise causal association, cell type-specific enrichment analysis could offer potential insights at the cellular level into the etiology of hypothyroidism from a distinct perspective compared to MR. We consider it as an important complementary cellular feature of hypothyroidism alongside MR results. The interpretation of the conclusions regarding the cell type specific enrichment analysis have been further modified in the results section.

Lines 270-282: âIn our MR analysis, the GWAS of immune cell was conducted based on peripheral blood samples. However, there is currently a lack of available GWAS data for immune cells at the tissue level (i.e. thyroid tissue). To understand the enrichment of cell types in thyroid tissues of hypothyroid population, a cell type-specific enrichment analysis (CSEA) was performed. Among the 12 cell types present in thyroid tissue, we identified eight cell types that exhibited nominally significant associations with at least one hypothyroidism dataset (Supplementary Data 11), which were endothelial cell, endothelial cell (APC), smooth muscle cell, T cell, macrophage, dendritic cell, B cell and B cell (Plasmacyte). The result of CSEA could be an important complementary cellular feature of hypothyroidism alongside MR resultsâ (Results section)

Lines 375-376. What are the âunknown metabolitesâ?

Response: Sorry for the potential confusion caused. The 1,400 plasma metabolites included 1,180 annotated metabolites and 220 unannotated metabolites. âunknown metabolitesâ indicated 220 unannotated metabolites in the previous version. In order to mitigate potential confusion for the reader, we have made a modification in the revised version by directly replacing the initial character of â1,400 metabolitesâ with the character of â1,180 annotated metabolitesâ. Additionally, we have omitted the previous statement of âunknown metabolitesâ.

Lines 541-543: âThe GWAS data for 1,180 annotated plasma metabolites were extracted from a metabolomics study in the Canadian Longitudinal Study on Aging (CLSA) cohort 29â (Methods section)

Lines 117-118. Did the authors mean that F statistic of exposure ~ genetic variant should exceed 10? It does not make sense otherwise. Also, to which F-statistic does the F column refer to in the supplementary tables?

Response: That is indeed our intention. An F statistic of â¥10 indicates a relatively low risk of weak instrument bias in MR analysis (Palmer TM, et al, Stat Methods Med Res, 2012;21(3):223-242). The âFâ column in the supplement data is indeed an âF-statisticâ. We've changed it in the supplementary data.

Lines 161-163: âThe F statistic of the utilized IVs exceeded 10 for all exposure data, which indicates a relatively low risk of weak instrument bias (Supplementary Data 2 and 3)â (Results section)

Lines 593-594: âF statistic of â¥ 10 indicates the absence of significant weak instrumental biasâ (Methods section)

According to the study's results, TNFSF11 acts as a protective factor for hypothyroidism. However, the proposed drugs (lines 235-236) act as inhibitors of the RANK/RANKL system. This appears to be a contradiction.

Response: We agree with the contradiction you raised. Given that we removed the inflammatory protein in the revised version, the drug reutilization section on TNFSF11 was removed accordingly. Thus, we did not discuss this in the revised version.

Reviewer #2:

This study aimed to assess the (causal) relationship of immunological and inflammatory proteins with hypothyroidism. Authors analysed several data and used a range of well-known statistical genetic methods. To me, the subject of this study is relevant and important. Also, the methods used are well-regarded, and overall, the analysis appeared to have been rigorously conducted. However, it occurs to me that the study has not been well presented coupled with several issues around grammatical errors. This study may benefit from professional language editing. Here are some specific comments for authors to consider in improving the quality and presentation of their study.

Response: We thank you for your helpful comments. In our revised edition, we have adopted your suggestions and made some corrections to the description and grammar of the manuscript. Please see our responses to your specific comments below:

Abstract
1. Lines 46 â 48 which were meant to set the tone of the study were not well constructed. Authors aimed to provide a rationale for their study, but this did not come across clearly. Please, rewrite the sentence.

Response: Thank you for your suggestion. We have rewritten the corresponding sentence in the revised version.

Lines 46-48: âWhile circulating metabolites and immune system have been increasingly linked to hypothyroidism risk, the causality underlying these associations remains largely uninterrogatedâ (Abstract section)

Introduction
1. I will recommend a revision of this introduction section, with a particular focus on clarity of expression, and grammatical accuracy. Authors need to rewrite a substantial part of this introduction to meet publication standards. There may be a need for professional English language editing to improve the presentation of this work. For example, âGiven that the heterogeneity of the elderly population, it is imperative to formulate new preventive strategies against the diseaseâ (lines 69 â 71) is not clear. The same comment applies to lines 72 â 75, line 76 (âinvestigatedâ), line 86 (âlower susceptibilityâ), etc. In line 87, âsignificantâ should be used in relation to statistical assessment.

Response: Thank you for your suggestion. The introduction section has undergone substantial revisions in terms of expression and grammar to adhere to the publication standards. In the revised version, the sentence you pointed out in lines 69-71, 72-75 and 76 were deleted. The other sentences you pointed out have been rewritten.

Lines 107-108: âMendelian randomization (MR) is a statistical method for inferring causality, which is not susceptible to reverse causality and confounding factors 8â (Introduction section)

Lines 111-113: âIn recent years, omics-based MR, has gained substantial progress in the 9,10
identification of disease targets, biomarkers, and potential pathogenic pathways â (Introduction section)

2. Authors made an effort to present their study, however, they have not provided sufficient justification or a convincing rationale for it. This issue may be related to the clarity of writing, which the authors can work on improving. In my opinion, addressing this matter is crucial for the paper to progress.

Response: Thank you for pointing out the major issue of our research. In the revised version, we have striven to make changes to the organization of the language to meet the standards for publication.

3. Lines 89 â 90, several studies have been conducted on this topic. What is the summary of the findings of those studies and how does the current study contribute or add something new? In other words, what are the limitations or gaps in knowledge that your study seeks to fill?

Response: Thank you for your suggestion. We have rewritten the corresponding sentences you pointed out.

Lines 121-127: âA previous MR study investigated the causal relationship between 69 metabolites and autoimmune hypothyroidism 11. They found that N-(3-furoyl) glycine, pipecolate, phenylalanine, allantoin, indololactate and alanine were associated with autoimmune hypothyroidism. However, the study encompassed a limited number of metabolites and lacked a validation cohort for hypothyroidism, which affected the robustness of the findings to some extentâ (Introduction section)

Lines 128-135: âTo address these limitations, we meticulously assembled the most comprehensive GWAS data available to date for plasma metabolomics, immune cell traits and two large-scale hypothyroidism cohorts. Leveraging these datasets, we performed a two-sample MR study to systematically investigate the potential causal associations of metabolites and immune system with hypothyroidismâ (Introduction section)

Results

1. It is not clear why only very few instruments were used for the Mendelian randomisation analysis, for example, for the IVW model (Supplementary Table S9, for example). How reliable are the results based on these limited instruments? If there are not enough genome-wide significant SNPs, could authors have relaxed the cut-off point to say, genome-wide suggestive level?

Response: Thank you for your suggestion. As you suggested, we adjusted the p-value threshold of instrumental variable selection from 5e-8 to 5e-6 and excluded those exposures with less than three SNPs. Due to the adjusted p-value threshold, there are sufficient IVs (nâ¥3) used for the pleiotropy test, including Egger intercept test and MR-PRESSO global test. We believe these efforts will reduce the threat of pleiotropy to our findings.

Lines 157-163: âAt the LD clumping criteria of pairwise LD R2 < 0.001 within a 10,000 kb window and the significance threshold of P-value < 5Ã10 -6, nearly 96% of exposures (1,150 exposures for plasma metabolites and 683 exposures for immune cells traits) had at least three SNPs for subsequent analysis (Supplementary Data 2 and 3). The F statistic of the utilized IVs exceeded 10 for all exposure data, which indicates a relatively low risk of weak instrument bias (Supplementary Data 2 and 3)â (Results section)

Lines 572-576: âFor each exposure, we initially filtered SNPs with a threshold of P-value < 5Ã10-6 and minor allele frequency (MAF) > 0.05. Subsequently, we performed clumping by the PLINK software using a cut-off of R2 < 0.001 and window size of 10,000 kb, utilizing the 503 European samples data from the 1000 Genomes project as our reference panel 34â (Methods section)
Lines 580-582: âWe also systematically removed exposures that had fewer than three IVs, thereby ensuring the conduction of subsequent pleiotropy testâ (Methods section)

2. Legend on Figure 2C is somewhat confusing, as it came across as if PH4 = 0.8 represents no support. Please, consider a correction here. This comment also applies to Figure 3B.

Response: Sorry for the potential confusion caused. We have changed this in the new Figure 4.

3. Lines 180 â 181, do authors mean the TNFSF11 level was causally protective for hypothyroidism?

Response: Sorry for the potential confusion caused. That is indeed our intention. In the revised version, we excluded GWAS data associated with inflammatory proteins in exposure for two primary reasons: 1) limited number of instrumental variables (n<3), despite relaxing the threshold, which hampers subsequent sensitivity analyses and

pleiotropy tests; 2) our original findings did not reveal any novel causal associations between inflammatory proteins and hypothyroidism. For instance, TNFSF11 (also called RANKL) mentioned in this question has been previously reported in other MR studies and thus cannot be considered as a new finding (Yang H, et al, J Clin Endocrinol Metab, 2023;108(2):433-442). Considering the above reasons, we only focused on the causal effects of metabolites and immune cells on hypothyroidism in the revised version.

4. Line 209, .â.. still remain to be fully explored.â Are there studies that have looked at them? if yes, cite them appropriately, and show what the current study is adding.

Response: Sorry for the potential confusion caused. This is actually a misunderstanding of our writing. We have changed the relevant statement.

Lines 332-335: âPrevious studies have reported a series of complications associated with hypothyroidism, including diabetes mellitus in pregnancy 12, diabetic nephropathy 13, heart failure 14, chronic kidney disease 15. An assessment of the shared metabolic and immune cellular mechanisms underlying hypothyroidism and its complications is crucial for gaining a comprehensive understanding of the underlying pathophysiology and potentially managing disease progression. Therefore, we explored the association between putatively causal traits and the above-mentioned complicationsâ (Results section)

Discussion
The discussion section is also not well presented. Authors need to do a bit of more work here to improve this section. Please, check how previous studies discuss their findings. Some specific points here are listed below.
1. Line 244, ââ¦this work is among the first â¦â is rather ambiguous. If it is the first, say so, If there are previous studies, authors need to acknowledge, cite them appropriately and state how their work advances those.

Response: Sorry for the potential confusion caused. We have modified related sentences in the revised version.

Lines 403-416: âAlthough the majority of the 20 putatively causal traits were first reported by our study. Several of them are consistent with the findings reported in previous research. The positive association between sphingomyelin (d18:1/20:0, d16:1/22:0) and hypothyroidism risk is consistent with a metabolomic analysis involving 27 cases in the Hashimotoâs thyroiditis with subclinical hypothyroidism (HTS) group. They found that the HTS group exhibited elevated levels of sphingomyelin compared to the control group 16. The 1-(1-enyl-palmitoyl)-2-oleoyl-GPE (p-16:0/18:1), a derivative of glycerophospholipids (GPL), was found to be associated with decreased hypothyroidism risk in our study. This metabolite in the hypothyroidism group depicted a decreasing trend via UPLCâQ-TOF/MS in a study enrolled discovery and validation sets of 175 and 300 participants, respectively 17. All of these showcase the validity of our findingsâ (Discussion section)

2. Line 248 âGPLâ. Authors used this term severally without first defining it. Please, define all terms, abbreviations, etc, at first use.

Response: We have modified the term.

Lines 411-413: âThe 1-(1-enyl-palmitoyl)-2-oleoyl-GPE (p-16:0/18:1), a derivative of glycerophospholipids (GPL), was found to be associated with decreased hypothyroidism risk in our studyâ (Discussion section)

3. Line 250 âThis result was confirmed through â¦â Is it confirmed by the study, or your findings confirmed the study, or your findings align/agree with ...?

Response: Sorry for the potential confusion caused. We have modified related sentences in the revised version.

Lines 405-409: âThe positive association between sphingomyelin (d18:1/20:0, d16:1/22:0) and hypothyroidism risk is consistent with a metabolomic analysis involving 27 cases in the Hashimotoâs thyroiditis with subclinical hypothyroidism (HTS) groupâ (Discussion section)

4. Grammar/punctuation errors, line 262 (ârole in both Theseâ)

Response: We have modified related sentences in the revised version.

Line 416: âAll of these showcase the validity of our findingsâ (Discussion section)
5. I was expecting information on how the identified âtargetsâ could be repurposed (noted by authors in lines 232 â 241) in the discussion section.

Response: Thank you for your interest in drug reutilization. We have added related statements in the discussion section.

Lines 471-483: âOur study also provides insights into drug reutilization for hypothyroidism. Both CD3 antigen and Sphingomyelin (d18:1/20:0, d16:1/22:0) were identified as risk factors for hypothyroidism in our study. Glofitamab targeting CD3 have been developed to treat relapsed or refractory diffuse large B-cell lymphoma. Glofitamab is a bi-specific monoclonal antibody that specifically targets the CD3 protein complex on T cells. By binding to CD3, Glofitamab facilitates the formation of immune synapses, which recruit T cells to CD20-expressing B cells and ultimately resulting in the elimination of target B cells 26. Therefore, Glofitamab could be a candidate novel drug candidate for hypothyroidism due to its ability to target lytic B cells. As a recombinant human acid sphingomyelinase, Olipudase alfa exhibits the capability to hydrolyze sphingomyelin 27, rendering it a promising therapeutic candidate for the management and prevention of hypothyroidismâ (Discussion section)

Methods
- Lines 351 â 354, Authors chose the GWAS data with a lower sample size as the discovery set, I have no problem with this, but any justification for such a decision? Also, is it possible to put all the data descriptions under one heading, say âGWAS data sourceâ rather than spreading them out under different sub-headings?

Response: Sorry for the potential confusion caused. In our original version, the GWAS for the finngen R9 study was conducted relatively new compared to another GWAS dataset, and the finngen R9 study used the mixed model logistic regression method (SAIGE) for association analysis. SAIGE can effectively control case-control imbalance and sample correlation in large-scale genetic association studies (Zhou W, et al, Nat Genet, 2018;50(9):1335-1341). This was our intention in choosing the finngen R9 study as the discovery dataset.

It is important to emphasize that, we identified a serious issue pertaining to the selection of replicated datasets in previous version. Our replication dataset is derived from published literature and is actually a meta-analysis of GWAS data from the finngen study and the UK biobank (Mathieu S, et al, iScience, 2022;25(9):104992). This means that the replication dataset is not independent due to population overlap with the discovery dataset. Therefore, in the revised version, we re-selected a replication dataset (14,871 cases and 391,429 controls) that included only the UK biobank population. Considering that the finngen R9 project (40,926 cases and 274,069 controls) has a larger number and proportion of cases, we still chose it as the discovery dataset for hypothyroidism.

Lines 552-557: âData on the associations with hypothyroidism were obtained from FinnGen R9 study 31 and UK biobank 32. We treated the FinnGen R9 study as the discovery study, which comprised 40,926 cases and 274,069 controls. The other study was considered as the replication study, which comprised 14,871 cases and 406,300 controlsâ (Methods section)

We have consolidated the relevant descriptions of GWAS sources under one heading.

Line 539: âGWAS data sourceâ (Methods section)

- Lines 376 â 377, âwe additionally excluded those unknown metabolites to obtain more definitive association results.â Is not clear.

Response: Sorry for the potential confusion caused. The 1,400 plasma metabolites

included 1,180 annotated metabolites and 220 unannotated metabolites. âunknown metabolitesâ indicated 220 unannotated metabolites in the previous version. In order to mitigate potential confusion for the reader, we have made a modification in the revised version by directly replacing the initial character of â1,400 metabolitesâ with the character of â1,180 annotated metabolitesâ. Additionally, we have omitted the previous statement of âunknown metabolitesâ.

Lines 541-545: âThe GWAS data for 1,180 annotated plasma metabolites were extracted from a metabolomics study in the Canadian Longitudinal Study on Aging (CLSA) cohort 29. This study focused on 8,299 unrelated European ancestry individuals in CLSA who had circulating plasma metabolites measured by Metabolon HD4 platformâ (Methods section)

- In line 399, âMR requirementsâ should be âMR assumptionâ.

Response: Thank you for your suggestion. We have revised the relevant statement.

Lines 606-608: âOR or Î² value with 95% confidence interval were evaluated to identify causal effects between exposures and outcomes, assuming MR assumptions were metâ (Methods section)

Reviewer #3:

Zheng and colleagues apply mendelian randomization and colocalization analysis to identify ten putative causal traits of hypothyroidism by leveraging multi-omics data including inflammatory proteins, plasma metabolites and immune cell traits. Additionally, the following-up mediation analysis reveal the potential mediating role of thyroid-stimulating hormone and the cell type-specific enrichment analysis demonstrate the important role of lymphocytes in the pathogenesis of hypothyroidism. Overall, this manuscript is well-organized, offering a clear and logical structure that enhances understanding. However, I also found multiple areas for improvement in the manuscript, including (a) additional results and interpretation of the sensitivity in the results presented, (b) additional clarity in the methods presented, (3) additional figures for the colocalization analysis and (4) lack of analysis in the non-European population.

Response: we thank you for your positive comments. We have modified the four important points you raised in the revised version. Please see our responses to your specific comments below:

I outline my detailed comments below.

1. Abstract: I suggest that the authors could incorporate some information regarding the identified putative causal traits within the text of the Abstract since these putative causal traits are the most important findings in this manuscript.

Response: Thank you for the suggestion. We have revised the abstract section to include key findings on cellular and metabolic characteristics causally associated with hypothyroidism.

Lines 55-67: âBy combining MR results from two large-scale cohorts, we ultimately identified 20 putatively causal traits, including five plasma metabolites and 15 immune cell traits. CD3 on CD28+ CD4+ T cell and 1-(1-enyl-palmitoyl)-2-oleoyl-GPE (p-16:0/18:1) demonstrated the most pronounced positive and negative associations with hypothyroidism risk, respectively. The odds ratio and 95% confidence interval were 1.09 (1.07, 1.12) and 0.81 (0.75â0.87), respectively. None of these 20 associations was biased by weak instruments, horizontal pleiotropy, or reverse causationâ (Abstract section)

2. Results - Mendelian Randomization: I suggest that the authors could incorporate some results and interpretations of sensitivity analysis after each Mendelian Randomization, for example, Cochranâs Q statistic, MR-Egger regression and MR-PRESSO, which are the

methods that authors pointed out at the Methods section and supplementary data.

Response: Thank you for the suggestion. We have added related results about sensitivity analysis in the revised version.

Lines 180-186: âAmong them, the association between kynurenine and hypothyroidism was not passed the sensitivity analysis due to the inconsistent effect direction of the additional method (MR Egger) (Supplementary Data 4). Other 16 significant associations were not biased by horizontal pleiotropy (Egger P for intercept > 0.05 and MR-PRESSO global test P > 0.05), heterogeneity among IVs (P for Cochranâs Q > 0.05) or reverse causation (P for MR Steiger test > 0.05) (Supplementary Data 4)â (Results section)
Lines 233-236: âAmong them, two associations of TCRgd T cell to lymphocyte ratio and CD28- CD127- CD25++ CD8 + T cell to T cell ratio with hypothyroidism were not passed the sensitivity analysis due to the inconsistent effect direction of the additional method (MR Egger) (Supplementary Data 7)â (Results section)

Lines 347-351: âThree associations among them were not passed the sensitivity analysis due to the inconsistent effect direction of the additional method (MR Egger), which were 1-(1-enyl-palmitoyl)-2-oleoyl-GPE (p-16:0/18:1)-chronic kidney disease, CD27 on IgD- CD38+ B cell-heart failure and CD3 on CD39+ CD4+ T cell-diabetic nephropathy pairs (Supplementary Data 13)â (Results section)

3. Results â P value threshold: in the section of âAssociations between plasma metabolites and hypothyroidismâ, the authors used FDR as the multiple testing correction, while in the section of âAssociation between immune cell traits with hypothyroidismâ, the authors used Bonferroni method as the multiple testing correction. Could the authors provide with the reasons using two different multiple testing correction methods?

Response: Sorry for the potential confusion caused. In the previous version, we adjusted for multiple tests in two methods, including Bonferroni method and FDR under Benjaminiâ Hochberg method, which could potentially lead to confusion as you mentioned. In the revised version, we consistently used the Bonferroni method for multiple test correction across all analyses.

Lines 170-174: âAfter adjustment for multiple testing by Bonferroni method (P-value < 4.35Ã10-5), we identified that 17 plasma metabolites were significantly associated with hypothyroidism risk using inverse-variance weighted method (Fig. 2a and Supplementary Data 4)â (Results section)

Lines 226-230: âAfter adjustment for multiple testing by Bonferroni method (P-value < 7.32Ã10-5), we identified a total of 26 immune cell traits that were significantly associated with hypothyroidism using inverse-variance weighted method (Fig. 3a and Supplementary Data 7)â (Results section)

Lines 608-612: âAdjustment for multiple testing was conducted to avoid the potential influence of false positive findings. Bonferroni method was used across all MR analysesâ (Methods section)

4. Methods/Results â Colocalization analysis: I suggest that the authors could incorporate the statistical fine-mapping (SuSiE) into the colocalization analysis, which can relax the single causal variants assumption. [Wallace C. A more accurate method for colocalisation analysis allowing for multiple causal variants. PLoS Genet. 2021 Sep 29;17(9):e1009440. doi: 10.1371/journal.pgen.1009440. PMID: 34587156; PMCID: PMC8504726.]
Response: Thank you for the suggestion. We have added SuSiE method for colocalization analysis.

Lines 258-262: âAmong the 20 identified associations (five for metabolites and 15 for immune cell traits), nine (45%) of these 20 putatively causal trait-hypothyroidism pairs showed a high colocalization with (PP4 > 0.8) using traditional colocalization method and sum of single effects (SuSiE) method (Fig. 4 and Supplementary Data 10)â (Results section)

Lines 674-678: âGiven the limitation of traditional colocalization method in detecting exposure and outcome traits with multiple causal genetic variants, we employed the SuSiE method along with a genetic correlation matrix reference panel derived from individuals of European ancestry in Phase III of the 1000 Genome Project to identify multiple causal variants.â (Methods section)

5. Methods â Study design: I am curious about is there any specific reasons that the authors used FinnGen R9 study as the discovery study, while use another data as the replication study. Based on the sample size the authors showed at the Method section, the second data has more cases and controls compared to the FinnGen R9 study. Additionally, I also suggest that the authors may conduct the meta-analysis by combining these two studies to enhance the power for the following analyses.

Response: Sorry for the potential confusion caused. In our original version, the GWAS for the finngen R9 project was conducted relatively new compared to another GWAS dataset, and the finngen R9 project used the mixed model logistic regression method (SAIGE) for association analysis. SAIGE can effectively control case-control imbalance and sample correlation in large-scale genetic association studies (Zhou W, et al, Nat Genet, 2018;50(9):1335-1341). This was our intention in choosing the finngen R9 project as the discovery dataset.

It is important to emphasize that, we identified a serious issue pertaining to the selection of replicated datasets in previous version. Our replication dataset is derived from published literature and is actually a meta-analysis of GWAS data from the finngen Project and the UK biobank (Mathieu S, et al, iScience, 2022;25(9):104992). This means that the replication dataset is not independent due to population overlap with the discovery dataset. Therefore, in the revised version, we re-selected a replication dataset (14,871 cases and 391,429 controls) that included only the UK biobank population. Considering that the finngen R9 project (40,926 cases and 274,069 controls) has a larger number and proportion of cases, we still chose it as the discovery dataset for hypothyroidism.

Lines 552-557: âData on the associations with hypothyroidism were obtained from FinnGen R9 study 31 and UK biobank 32. We treated the FinnGen R9 study as the discovery study, which comprised 40,926 cases and 274,069 controls. The other study was considered as the replication study, which comprised 14,871 cases and 406,300 controlsâ (Method section)
We appreciate the meta-analysis method you provided. In our modified version, meta-analysis was applied in two scenarios. 1) We performed a meta-analysis of the MR results of the discovery cohort and the replication datasets. 2) Prior to the colocalization analysis, we performed a meta-analysis of the discovery and replication datasets using the fixed effects model weighted by standard error.

Lines 616-618: âMeta-analyses of MR result in both the discovery and replication cohorts were calculated using the random effect model (R package meta, version 7.0-0)â (Method section)

Lines 654-658: âTo increase power in colocalization analysis, we used the METAL software to meta-analyze the summary statistics of hypothyroidism derived from FinnGen R9 study and UK biobank 47. The analysis was conducted using a fixed-effects model and the associations from two data sources were weighted by standard error of GWAS estimatesâ (Method section)

6. Methods â Colocalization analysis: I suggest that the authors could clarify how they set up the window size/region to select the variants for each colocalization analysis.

Response: Thank you for the suggestion. For each exposure, we selected regions that had 500 kb windows upstream and downstream of each instrumental variable in MR for analysis, and the instrumental variable with the largest posterior probability for H4 (PP4) was reported. (Method section)

Lines 678-682: âFor each exposure, we selected regions that had 500 kb windows upstream and downstream of each instrumental variable in MR for analysis 49, and the

instrumental variable with the largest posterior probability for H4 (PH4) was reportedâ (Method section)

7. Methods/data sources: The data sources that the authors used for Mendelian Randomization and Colocalization analysis were from European populations, which might limit the reliability when extrapolating the findings to non-European populations. I recommend that the authors could use available non-European data sources.

Response: Thank you for the suggestion. Unfortunately, based on available non-European GWAS data, we did not find GWAS data for the putative causal metabolites and immune cell traits identified in our study. We were aware that this would have implications for the generalization of our findings to populations outside of Europe. We have described this limitation in the discussion section.

Lines 518-520: âSecond, the scope of our analysis was limited to Europeans, thereby constraining the generalizability of our findings to other populationsâ (Discussion section)

8. Methods/data sources: The data sources that the authors used for Mendelian Randomization and Colocalization analysis were from European populations, which might limit the reliability when extrapolating the findings to non-European populations. I recommend that the authors could use available non-European data sources.

Response: Thank you for the suggestion. Unfortunately, based on available non-European GWAS data, we did not find GWAS data for the putative causal metabolites and immune cell traits identified in our study. We were aware that this would have implications for the generalization of our findings to populations outside of Europe. We have described this limitation in the discussion section.

Lines 518-520: âSecond, the scope of our analysis was limited to Europeans, thereby constraining the generalizability of our findings to other populationsâ (Discussion section)

9. Figure2c/3c â Colocalization analysis: I suggest that the authors could generate the colocalization plot for the colocalization analysis, which is helpful for the interpretation of the colocalization result. Here is R function to create the coloc plot: locuscompare(). https://hanruizhang.github.io/GWAS-eQTL-Colocalization/

Response: Thank you for the suggestion and code. We have used âlocuscompare ()â function to generate a colocalization plot in new Supplementary Figure 1.

Lines 683-684: âLocuscomparer package (1.0.0) was used to visualize the results of colocalizationâ (Method section)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors addressed every point of the precedent review. The quality of the study was greatly improved in the revised version and the authors' findings, although less extensive, are more robust. This being said, I still have a few comments that I think should be addressed before publication.

Line 58. You cannot say that horizontal pleiotropy does not play a role when testing with Egger intercept test. (Intercept not equal to 0 -> horizontal pleiotropy, but the reciprocal is not true.) What you can say is something like "no evidence of horizontal pleiotropy was found (from your regression-based method)" or "all finding showed balanced pleiotropy".

Line 167. Given that all CD3 related traits show similar effects, it would be interesting to highlight shared IVs between these putatively causal associations (or LD correlations between theses IVs) to answer this question: "Are all of those traits causal or some signals result from shared genetic determinants between these biologically similar exposures?" In other words, are these signals genetically independent from each other or pleiotropy could play a role? An approach using Multi variable MR could give further insights to that matter. This comment is related to the precedent: if two traits share a very large proportion of their genetic determinants, the Egger Intercept method will not catch potential cases of horizontal pleiotropy since asymptotically all instruments are associated to both traits.

Lines 245-247. “Although the majority of the 20 putatively causal traits were first reported by our study. Several of them are consistent with the findings reported in previous research.” Should this be a single sentence?

Lines 363-366. According to the provided reference the value of 2B^2*MAF*(1-MAF) corresponds to the contribution of the SNP to the genetic variation of the trait and not the proportion of the overall variance of the trait explained by the SNP. From my understanding, these are different concepts. The reference 37 does not provide a motivation for the approximation of R2 by this value (which they call “effect size” or ES). If you do not dispose of the R2 in the summary statistics of exposure data and cannot provide evidence that ES approximates R2, I would suggest you evaluate the F-statistic with F = B^2/SE^2 where B is the estimated slope of the association of the IV to the exposure and SE is the standard error of B (PMID: 31924771).

Lines 372-373. Now that all MRs have more than 2 IVs, this parenthesis is useless and could be confusing. It could be simply removed.

Reviewer #2

(Remarks to the Author)
Authors have satisfactorily addressed all my concerns. I have no further comments.

Thanks

Reviewer #3

(Remarks to the Author)
A well-written manuscript. All concerns from the first review have been addressed. No further comments at this time.

Author Rebuttal letter:

Dear reviewers,

RE COMMSBIO-24-0721A

Multi-omics insight into the molecular and cellular characteristics in the pathogenesis of hypothyroidism

We would like to thank you for your helpful comments and suggestions for improving our manuscript. We have uploaded a clean version of the revised manuscript and a version with changes in red color. Our point-by-point responses to your comments are presented below in blue color.

Reviewer #1 (Remarks to the Author):

The authors addressed every point of the precedent review. The quality of the study was greatly improved in the revised version and the authorsâ findings, although less extensive, are more robust. This being said, I still have a few comments that I think should be addressed before publication.

Response: Thank you for your positive comments on our revised version. We have responded to your questions in the new revised version. Please see our responses to your specific comments below:

Line 58. You cannot say that horizontal pleiotropy does not play a role when testing with Egger intercept test. (Intercept not equal to 0 -> horizontal pleiotropy, but the reciprocal is not true.) What you can say is something like âno evidence of horizontal pleiotropy was found (from your regression-based method)â or âall finding showed balanced pleiotropyâ.

Response: Thank you for the suggestion. We have modified the sentence you mentioned.

Lines 58-60: âNo evidence of horizontal pleiotropy, heterogeneity among instrumental variables or reverse causation were found for these 21 significant associations.â (Abstract section)

Line 167. Given that all CD3 related traits show similar effects, it would be interesting to highlight shared IVs between these putatively causal associations (or LD correlations between theses IVs) to answer this question: âAre all of those traits causal or some signals result from shared genetic determinants between these biologically similar exposures?â In other words, are these signals genetically independent from each other or pleiotropy could play a role? An approach using Multi variable MR could give further insights to that matter.

This comment is related to the precedent: if two traits share a very large proportion of their genetic determinants, the Egger Intercept method will not catch potential cases of horizontal pleiotropy since asymptotically all instruments are associated to both traits.
Response: Thank you for pointing out this interesting question. Eight CD3 related immune traits were identified as potential causal traits for hypothyroidism in our study and show similar effect. These eight traits all belong to CD3 antigens on the surface of certain T cell subtypes, so it is well understood that they may share genetic factors or correlated genetically. It is noteworthy that the GWAS analyses of these immune traits were based on high-throughput flow cytometry sorting, and a specific feature of high-throughput experiments is the unique pattern of correlation between candidate risk factors formed by potential biological processes.

To better answer the question you mentioned, we performed a MR based on Bayesian model averaging (MR-BMA), which is an extension of multivariate MR (MVMR), to identified real causal traits (Zuber V, et al, Nat Commun, 2020;11(1):29). Traditional MVMR is designed for a small number of risk factors and cannot be scaled up to the dimensions of high-throughput experiments. Compared with traditional MVMR, MR-BMA can prioritize and select causal risk factors in a Bayesian framework from a set of high-dimensional relevant candidate risk factors. In other words, MR-BMA can be used to determine which of a set of related risk factors with common genetic predictors are causal drivers of disease risk. Considering the large number of CD3 related immune traits, the correlation among these immune traits, and the unique correlation pattern of high-throughput experiments, we chose MR-BMA rather than traditional MVMR for analysis.

In short, taking into account all the exposures specified, MR-BMA iterates over a number of potentially "real" causal models. For each exposure, a Marginal inclusion probability (MIP) is calculated, representing the posterior probability that the exposure factor x appears in a true causal model for a given z iteration. The method also estimates the Model-averaged causal effect (MACE), which represents the exposureâs weighted average direct causal effect on risk across all candidate models. It is worth noting that MACE should not be interpreted in absolute terms. It will be biased towards zero due to the contraction applied in variable selection. But this metric can be used to understand the direction and magnitude of the impact relative to other exposures included in the model.

According to the previous study (Lord J, et al, Proc Natl Acad Sci U S A, 2021;118(16):e2009808118)ï¼we set z to 10,000, the prior probability to 0.1, and prior variance (Ï2) to 0.25. MR-BMA result showed CD3 on HLA DR+ CD4+ T cell was the most likely potential causal trait (MIP=0.424, MACE=0.018, P-value=7.9Ã10-3, Ranking=1). We also adjusted the prior probability to 0.5 for the sensitivity analysis, and the result is consistent with the previous conclusion (MIP=0.396, MACE=0.016, P-value=2.9Ã10-2, Ranking=1).

We have added the relevant results in Supplementary Data 11 and added relevant content in the manuscript. Thank you again for this valuable question.

Lines 98-101: âSecond, MR based on Bayesian model averaging (MR-BMA) was performed to determine which of a set of related risk factors with common genetic predictors were causal drivers of hypothyroidism risk.â (Introduction section)

Lines 119-122: âBesides, we performed a series of follow-up analyses, including colocalization analysis, MR-BMA analysis, cell type-specific enrichment analysis, hypothyroidism-related complications association analysis and druggability analysis.â (Results section)

Lines 201-213: âInterestingly, eight CD3 related immune traits were identified as potential causal traits for hypothyroidism in our study and show similar effect. To determined which of a set of these immune traits with common genetic predictors were causal drivers of hypothyroidism risk, we employed MR-BMA analysis, which is an extension of multivariate MR (MVMR) 13. According to the previous study 14ï¼we set z to 10,000, the prior probability to 0.1, and prior variance (Ï2) to 0.25. Marginal inclusion probability (MIP) and Model-averaged causal effect (MACE) were calculated. MR-BMA result showed CD3 on HLA DR+ CD4+ T cell was the most likely potential causal trait (MIP=0.424, MACE=0.018, P-value=7.9Ã10-3, Ranking=1) (Supplementary Data 11). We also adjusted the prior probability to 0.5 as the sensitivity analysis, and the result is consistent with the previous conclusion (MIP=0.396, MACE=0.016, P-value=2.9Ã10-2, Ranking=1) (Supplementary Data 11)â (Results section)

Lines 474-491: âIn short, taking into account all the exposures specified, MR-BMA iterates over a number of potentially "real" causal models 13. For each exposure, a MIP value is

calculated, representing the posterior probability that the exposure factor x appears in a true causal model for a given z iteration. The method also estimates the MACE, which represents the exposureâs weighted average direct causal effect on risk across all candidate models. It is worth noting that MACE should not be interpreted in absolute terms. It will be biased towards zero due to the contraction applied in variable selection. But this metric can be used to understand the direction and magnitude of the impact relative to other exposures included in the model.

We combined the IVs from the eight CD3 related immune traits, and further clumped by linkage disequilibrium (R2 < 0.001 within a window of 10,000 kb) to ensure that the IVs were independent. SNP effects and corresponding standard errors were then extracted from the corresponding immune GWAS summary statistics and harmonized with the outcome hypothyroidism GWAS information. We set z to 10,000, the prior probability to 0.1, and prior variance (Ï2) to 0.25 in the main analysis. We additionally set prior probability to 0.5 as sensitivity analysis. The analysis was conducted using the mrbma package (0.1.0).â (Methods section)

Lines 559-561: â13. Zuber, V., Colijn, J. M., Klaver, C. & Burgess, S. Selecting likely causal risk factors from high-throughput experiments using multivariable Mendelian randomization. Nat Commun 11, 29 (2020).â (References section)

Lines 562-564 â14. Lord, J. et al. Mendelian randomization identifies blood metabolites previously linked to midlife cognition as causal candidates in Alzheimerâs disease. Proc Natl Acad Sci U S A 118, e2009808118 (2021).â (References section)

Lines 245-247. âAlthough the majority of the 20 putatively causal traits were first reported by our study. Several of them are consistent with the findings reported in previous research.â Should this be a single sentence?

Response: Thank you for finding the grammatical problem in the manuscript. We have rewritten the sentence in new revised version.

Lines 271-273: âAlthough the majority of the 21 putatively causal traits were first reported by our study, several of them are consistent with the findings reported in previous research.â (Discussion section)

Lines 363-366. According to the provided reference the value of 2B^2*MAF*(1-MAF) corresponds to the contribution of the SNP to the genetic variation of the trait and not the proportion of the overall variance of the trait explained by the SNP. From my understanding, these are different concepts. The reference 37 does not provide a motivation for the approximation of R2 by this value (which they call âeffect sizeâ or ES). If you do not dispose of the R2 in the summary statistics of exposure data and cannot provide evidence that ES approximates R2, I would suggest you evaluate the F-statistic with F = B^2/SE^2 where B is the estimated slope of the association of the IV to the exposure and SE is the standard error of B (PMID: 31924771).

Response: Thank you for the suggestion. We have re-calculated the F-statistic by the formula you offered. New F-statistic data was updated in the new Supplementary Data 2, 3, 4, 5, 7, 8 and 14.

Lines 387-390: âF statistic is calculated to test instrument strength through the formula F =Î²2/SE2, where Î² denotes the estimated genetic effect on exposure, SE represents standard error of Î² 38.â (Methods section)

Lines 627-630: â38. Richardson, T. G., Hemani, G., Gaunt, T. R., Relton, C. L. & Davey Smith, G. A transcriptome-wide Mendelian randomization study to uncover tissue-dependent regulatory mechanisms across the human phenome. Nat Commun 11, 185 (2020).â (References section)

Lines 372-373. Now that all MRs have more than 2 IVs, this parenthesis is useless and could be confusing. It could be simply removed.

Response: Thank you for the suggestion. We have modified the sentence in the new revised version.
Lines 399-401: âInverse-variance weighted (IVW) under a multiplicative random-effects model was used for main analysis methodâ (Methods section)

Reviewer #2 (Remarks to the Author):

Authors have satisfactorily addressed all my concerns. I have no further comments.

Thanks

Response: We sincerely appreciate all your efforts in reviewing the manuscripts and your suggestions have really improved the quality of our work!

Reviewer #3 (Remarks to the Author):

A well-written manuscript. All concerns from the first review have been addressed. No further comments at this time.

Response: We sincerely appreciate all your efforts in reviewing the manuscripts and your suggestions have really improved the quality of our work!

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
All my concerns were addressed by the authors. I have no further comments at this point. Good work!

Author Rebuttal letter:

Dear reviewers,

RE COMMSBIO-24-0721B

Multi-omics insight into the molecular and cellular characteristics in the pathogenesis of hypothyroidism

Reviewer #1 (Remarks to the Author):

All my concerns were addressed by the authors. I have no further comments at this point. Good work!

Response: We sincerely appreciate all your efforts in reviewing the manuscripts and your suggestions have really improved the quality of our work!

Zheng and colleagues apply mendelian randomization and colocalization analysis to identify ten putative causal traits of hypothyroidism by leveraging multi-omics data including inflammatory proteins, plasma metabolites and immune cell traits. Additionally, the following-up mediation analysis reveal the potential mediating role of thyroid-stimulating hormone and the cell type-specific enrichment analysis demonstrate the important role of lymphocytes in the pathogenesis of hypothyroidism. Overall, this manuscript is well-organized, offering a clear and logical structure that enhances understanding. However, I also found multiple areas for improvement in the manuscript, including (a) additional results and interpretation of the sensitivity in the results presented, (b) additional clarity in the methods presented, (3) additional figures for the colocalization analysis and (4) lack of analysis in the non-European population.

I outline my detailed comments below.

1. Abstract: I suggest that the authors could incorporate some information regarding the identified putative causal traits within the text of the Abstract since these putative causal traits are the most important findings in this manuscript.

2. Results - Mendelian Randomization: I suggest that the authors could incorporate some results and interpretations of sensitivity analysis after each Mendelian Randomization, for example, Cochran's Q statistic, MR-Egger regression and MR-PRESSO, which are the methods that authors pointed out at the Methods section and supplementary data.

3. Results – P value threshold: in the section of 'Associations between plasma metabolites and hypothyroidism ', the authors used FDR as the multiple testing correction, while in the section of 'Association between immune cell traits with hypothyroidism', the authors used Bonferroni method as the multiple testing correction. Could the authors provide with the reasons using two different multiple testing correction methods?

4. Methods/Results – Colocalization analysis: I suggest that the authors could incorporate the statistical fine-mapping (SuSiE) into the colocalization analysis, which can relax the single causal variants assumption. [Wallace C. A more accurate method for colocalisation analysis allowing for multiple causal variants. PLoS Genet. 2021 Sep 29;17(9):e1009440. doi: 10.1371/journal.pgen.1009440. PMID: 34587156; PMCID: PMC8504726.]

5. Methods – Study design: I am curious about is there any specific reasons that the authors used FinnGen R9 study as the discovery study, while use another data as the replication study. Based on the sample size the authors showed at the Method section, the second data has more cases and controls compared to the FinnGen R9 study. Additionally, I also suggest that the authors may conduct the meta-analysis by combining these two studies to enhance the power for the following analyses.

6. Methods – Colocalization analysis: I suggest that the authors could clarify how they set up the window size/region to select the variants for each colocalization analysis.

7. Methods/data sources: The data sources that the authors used for Mendelian Randomization and Colocalization analysis were from European populations, which might limit the reliability when extrapolating the findings to non-European populations. I recommend that the authors could use available non-European data sources.

8. Methods/data sources: The data sources that the authors used for Mendelian Randomization and Colocalization analysis were from European populations, which might limit the reliability when extrapolating the findings to non-European populations. I recommend that the authors could use available non-European data sources.

9. Figure2c/3c – Colocalization analysis: I suggest that the authors could generate the colocalization plot for the colocalization analysis, which is helpful for the interpretation of the colocalization result. Here is R function to create the coloc plot: locuscompare(). https://hanruizhang.github.io/GWAS-eQTL-Colocalization/