nature portfolio

Peer Review File

Transcript-aware rare genetic variant association analyses of cardiopulmonary traits in participants from the *All of Us* Research Program

Corresponding Author: Seung Hoan Choi

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
This manuscript proposes a transcript-specific masking strategy for rare variant association testing of exomic variants. The authors apply a burden test for pLOF and pDM variants across 118 phenotypes. The proposed approach aligns with established methodologies that test multiple masks and aggregate the resulting p-values using the Cauchy combination test, a well-accepted technique in gene-based analyses.

While the concept of a transcript-aware mask is a potentially valuable addition to the rare variant testing framework, the paper lacks methodological novelty and demonstrates only moderate significance. Furthermore, the absence of simulation studies limits the ability to evaluate whether the proposed method offers meaningful power improvements over existing approaches.

Overall, I am not convinced that the manuscript meets the threshold for publication in a high-profile journal.

1. As noted, the proposed approach is a straightforward application of a pLoF + pDM burden test, where burden scores are constructed for each transcript. The method then combines p-values across transcripts using the Cauchy combination test. This general framework—applying multiple variant masks and aggregating p-values via the Cauchy method—is well established in gene-based testing. Therefore, the manuscript does not present any clear methodological innovation.

2. The authors do not include any simulation studies to assess the potential power gains from using a transcript-aware mask. Such simulations would be valuable to clarify the conditions under which the proposed approach yields a meaningful improvement. For example, it would be helpful to understand how much the canonical and non-canonical transcripts need to differ in order to see a substantial power gain. Based on Figure 1, only TTN appears to show a notable improvement. Also In Table1, many of the novel associations show the most significant results for the canonical transcripts (“C-”).

3. Line 78: “The proposed framework demonstrated strong reliability, revealing 46 associations previously reported in rare variant studies.”
Since the pLoF+pDM burden test has already been applied in WES studies and is included as one of the candidate masks in the current framework, replicating known associations does not necessarily validate the reliability of the proposed method.

4. Line 180 paragraph: It is unclear what the authors intend to convey in this paragraph. Are they suggesting that transcript-aware masking improves cross-ancestry consistency? Or are they simply noting that the effects of pLoF+pDM variants are consistent across ancestry groups? If it is the latter, it is not evident how this observation is related to the proposed transcript-aware masking strategy.

Minor comments:
1. Figure 1 a: x-axis title: need to remove SNP?

(Remarks on code availability)
The code lacks both usage instructions and example datasets, making it difficult to reproduce or evaluate the results.

Reviewer #2

(Remarks to the Author)
The study from Zhang and colleagues proposes a transcript-aware burden test approach on 242,881 participants from the All of Us research program to assess the association with 129 cardiopulmonary traits. Genetic variants of interest included in the burden score are loss of functon (LoF) and predicted deleterious missense (pDM).
A key aspect of the work is the focus on transcripts, distinguishing canonical from non-canonical: the idea is simple, but important. Authors additionally created a “pseudo transcript”, which combines the most predicted deleterious variants observed in the canonical and non-canonical transcripts for a certain gene.
Methods are robust, data are well presented and treated according to rigorous quality control criteria. Overall, the study introduces a promising approach to better understand known associations and for discovering new ones in large heterogeneous cohorts.
I have a couple of comments for the Authors:

The logic behind the pseudo transcript creation is interesting, but I would add a cautionary note about the interpretation of the findings.
In figure 1b, pseudo transcripts associations typically coincide with the combined analysis, sometimes association is slightly improved, but sometimes is also diluted, like in TTN and cardiomyopathy. This may be due to the incorporation of variants less pathogenic in practice.
Indeed, despite the high confidence of the predictions, the actual impact of a genetic variant on a phenotype can be assessed only with in vitro experiments on appropriate models. This is important also for rare LoF, which sometimes are neutral, like the example of LRRK2 (Whiffin 2020 Nat Med). This also implies that variants, which are predicted to be less deleterious, could be pathogenic in fact. I would clarify this potential limitation.

Finally, because the study is of general interest, to facilitate the less experienced reader in understanding the rich pipeline and supplementary materials, I would move the extended data figure 1 - Study Flow Chart to the main text.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
In this study, the authors present a transcript-aware rare variant association test. The research question is relevant, particularly in the context of aggregation tests. However, I have significant concerns about the All of Us analyses, and some sections of the manuscript would benefit from greater clarity.
My main comment is about the genetic ancestry included in this study. If I understood correctly, the main analyses include all individuals, regardless of ancestry, except for the adjustment on common variants-based PCs. This choice is surprising given that population stratification is more pronounced for rare variants than for common variants, for which ancestry-aware meta-analyses are often performed. Even for the “ancestry-stratified” analyses, I have concerns about the justification of a “non-EUR” group.
Related to the previous point, why are there only 7 associations reported in Supp Fig 7 while 12 novel genetic associations are found? Is it because the other ones are not detected in EUR and non-EUR analyses?
The second main comment is about the additive value of looking at transcript-aware rare variant tests. Even if the benefit is highlighted in some examples, the overall gain is not clear. Looking at Figure 1, it seems for example that the signals identified in the combined analyses would also be detected in the proxy or canonical analysis. While the authors define 12 new signals, it is unclear whether this is due to the analysis of the All of Us dataset or to the transcript-aware analyses. It would be useful for example to give the percentage of genes detected only using the combined approach and not the canonical or pseudo transcripts.
Related to the previous point, the authors compare the p-values between the different approaches. I am not sure what the authors mean by “a 100% gain in significance”. Comparing the effect sizes of the tests might be more relevant.
In addition, I have additional comments:
- From what I understand, whole-exome sequencing data and not WGS data are analyzed. Please change accordingly or clarify.
- In Table 1, it would be useful to indicate how many transcripts are present in the combined test. Can the combined p-value be affected by the number of transcripts analyzed?
- Some supplementary materials (ex Supp Fig 9) are not referenced in the text
- Can the authors provide a list of all the phenotypes tested for association along with the same size? It is not so clear how are the clusters used, and how the phenotypes were selected.
- Were the analyses in UKB performed in a similar way as in All of Us?

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have addressed my previous comments and conducted additional simulation studies to evaluate the performance of the proposed method. The simulations generally demonstrate that the proposed approach adequately controls type I error rates. However, the power comparisons suggest that the T-aware approach is substantially more powerful than the two alternative methods (canonical and pseudo).

I am concerned that these results are somewhat misleading, as the simulation settings appear to favor scenarios in which the proposed method is expected to perform optimally. In the All of Us data analysis, however, the canonical and pseudo approaches often yielded smaller p-values than the proposed approach. According to the authors' response to Reviewer 2 (Comment 3), the T-aware method achieved the minimum p-value in only 31% of significant associations, and the total number of significant associations was not substantially different across methods.

To provide a more balanced and comprehensive evaluation, the authors should conduct power simulations under a broader range of scenarios. This would help clarify the conditions under which the proposed method achieves superior power, as well as situations in which its performance is comparable (e.g., via CCT) or potentially less favorable relative to existing approaches. Such analyses would offer a more realistic and informative assessment of the method's practical utility.

(Remarks on code availability)
The quality of the software package needs improvement, as it currently depends on a file path that is not accessible to external users. Specifically, the code includes a reference such as:
source('/UKBB_200KWES_CVD/GENESIS_adaptation_source.R')

The authors should revise the implementation to ensure that all required scripts and dependencies are properly included within the package structure, making it fully self-contained and reproducible for external users.

Reviewer #2

(Remarks to the Author)
I thank the Authors for considering my comments, and I have no additional remarks.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
I thank the authors for their responses, and I have no further comments.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors addressed my previous comments. I don't have any additional ones.

(Remarks on code availability)

Reviewer #1 (Remarks to the Author):

This manuscript proposes a transcript-specific masking strategy for rare variant association testing of exomic variants. The authors apply a burden test for pLOF and pDM variants across 118 phenotypes. The proposed approach aligns with established methodologies that test multiple masks and aggregate the resulting p-values using the Cauchy combination test, a well-accepted technique in gene-based analyses. While the concept of a transcript-aware mask is a potentially valuable addition to the rare variant testing framework, the paper lacks methodological novelty and demonstrates only moderate significance. Furthermore, the absence of simulation studies limits the ability to evaluate whether the proposed method offers meaningful power improvements over existing approaches.
Overall, I am not convinced that the manuscript meets the threshold for publication in a high-profile journal.

1. As noted, the proposed approach is a straightforward application of a pLoF + pDM burden test, where burden scores are constructed for each transcript. The method then combines p-values across transcripts using the Cauchy combination test. This general framework—applying multiple variant masks and aggregating p-values via the Cauchy method—is well established in gene-based testing. Therefore, the manuscript does not present any clear methodological innovation.

   Response: We thank the reviewer for pointing out the ambiguity of the statement. We agree that burden testing and the Cauchy combination test are established components of gene-based analysis. Our approach lies in introducing a framework that incorporates transcript-specific masking within rare variant association testing. This framework provides a biologically motivated refinement of multi-mask gene-based testing by integrating transcript structure, which is particularly relevant for genes with extensive isoform diversity. The key idea is to embed biologically meaningful transcript boundaries into standard rare variant frameworks rather than introducing a new statistical test. This framework is particularly relevant when a specific transcript is causally related to a trait, as illustrated for *TTN*. Our cardiopulmonary association analyses suggest the potential utility of this framework in such scenarios.

2. The authors do not include any simulation studies to assess the potential power gains from using a transcript-aware mask. Such simulations would be valuable to clarify the conditions under which the proposed approach yields a meaningful improvement. For example, it would be helpful to understand how much the canonical and non-canonical transcripts need to differ in order to see a substantial power gain. Based on Figure 1, only TTN appears to show a notable improvement. Also In Table1, many of the novel associations show the most significant results for the canonical transcripts ("C-").

   Response: Thank you for the helpful suggestions. We agree that several novel associations in Table 1 show the smallest p-value for the canonical transcript (C-). However, when the canonical transcript is the leading transcript, associations tend to involve fewer transcripts in total. Among the 5 novel associations in which the canonical transcript has the smallest p-value, two involve only 2 transcripts and two have 4 transcripts (Revised manuscript Table 1). The proposed framework shows greater benefit when many transcripts are involved (for example, *TTN* and *PPARG*) because it aggregates information across transcripts. When a few transcripts exist, the framework remains comparable to a canonical-only analysis because it also leverages the canonical transcript. Additionally, under the same QC, restricting to the canonical

transcript also led to fewer testable gene-phenotype pairs (1,739,173) than the transcript-aware and pseudo-transcript-only analyses (1,818,862). Restricting to cardiopulmonary traits may reduce the number of compelling examples (such as *TTN*) that demonstrate the benefit of the framework. We have acknowledged this limitation in the manuscript, but we still expect the framework to be broadly useful and to have great potential.

To provide a better illustration of the gains in power, we conducted extensive simulation studies with different scenarios to evaluate the performance of the proposed framework. In brief, we compared the performance of the proposed framework to several approaches (including standard method of using only canonical transcript) under varying proportions of causal variants assigned to the canonical transcript (30%-100%). The causal variant proportion controls how many true causal variants fall into the canonical transcript, and therefore how much the canonical transcript differs from the non-canonical transcripts. This helps illustrate when the proposed approach yields substantial power gains. Simulation results show that the transcript-aware method consistently achieved higher power than both the canonical-only and pseudo-transcript-only tests, particularly when the canonical transcript captured only part of the true causal signals. Therefore, in settings where a disease is caused by specific transcripts, our transcript-aware approach outperforms existing rare-variant framework. All results and details were updated to the revised manuscript ("Simulation" subsection in the **Results and Discussion** section, **Method** section, and **Supplementary** Figure 2, and 11, also showing below).

Manuscript Page 5-6, Lines 76-99:

Simulation Study

*Simulation results comparing four approaches, canonical-only, pseudo-only, T3-only (ideal case), and the proposed transcript-aware (T-aware) method, were summarized in **Fig.2**. All methods showed well-controlled type I error at α = 0.001 and α =0.01 (**Fig.2 a** and **c**). At α = 0.001, the proposed T-aware method kept the false positive rate close to the nominal level, while the T3-only approach was slightly lower than nominal (**Fig.2a**). At α = 0.01, all methods had type I error close to the nominal level (**Fig. 2c**) with overlapping confidence intervals (CIs). Similar patterns were observed at α = 0.05 (**Supplementary Fig.1a**), with the T3-only approach being slightly more conservative and T-aware being slightly above the nominal level. To further assess these behaviors, we performed additional simulations using a higher prevalence of 2% (**Supplementary Fig.2**). Consistent with the results at 0.2% prevalence, all methods controlled the type I error rate well at α = 0.001 and α = 0.01, with T3-only remaining slightly more conservative and T-aware slightly above the nominal level at α = 0.05. The pattern of the p-values at α = 0.05 was likely a result of both simulation noise from a finite number of replicates and the case–control imbalance presented in rare-variant analyses.*

*When comparing power, the T-aware method was more powerful than the canonical-only and pseudo-only approaches across all scenarios, with no overlap between the CIs for the estimated power (**Fig.2 b** and **d**). The difference in power was larger when the canonical transcript contained fewer causal variants. Notably, even when the canonical transcript contained all causal variants, the canonical-only approach was less powerful because it included many additional non-causal variants. As expected, the highest power was achieved when the most outcome-relevant transcript (T3) was known in advance and used alone. However, in reality, this true outcome-relevant transcript is generally unknown, so T-aware can serve as a practical surrogate for the ideal approach.*

Manuscript Page 17-18, Lines 279-302:

Simulation

We conducted simulations with n=200,000 individuals for a hypothetical gene. The gene had 300 rare and ultra-rare

genetic variants, of which 30 were truly causal and increased disease risk (same effect direction). All causal variants had minor allele frequencies (MAF) below 0.1%, with 80% ultra-rare (MAF ~ 0.0005%–0.005%) and 20% rare (MAF ~ 0.005%–0.05%). For simplicity, we assumed no linkage disequilibrium among these variants, which is reasonable for rare and ultra-rare variants. The variants were mapped to four transcripts (T1–T4), where T1 was the canonical transcript (longest) and T3 was the "best" transcript that contained all causal variants. Here T3 represented the most outcome-relevant transcript. The outcome was binary with a prevalence of 0.2%, roughly corresponding to a disease that occurs about 1 in 500 individuals, such as cardiomyopathy.

*We considered five simulation scenarios, each repeated 20,000 times. Scenario 1 was a null setting used to evaluate type I error, in which none of the transcripts contained causal variants. Scenarios 2–5 were used to assess power by varying the proportion of causal variants assigned to the canonical transcript, while transcript T3 always contained all causal variants. These proportions were 30%, 60%, 80%, and 100%, corresponding to 9, 18, 24, and 30 causal variants in the canonical transcript (**Supplementary Fig.11**). We compared four approaches (1) canonical-only, (2) pseudo-only, (3) T3-only, and (4) our proposed transcript-aware method, all using the burden test. Here canonical-only represented a standard approach that used a single transcript per gene without pre-selection. T3-only represented an idealized case in which the most informative transcript was known in advance. The "pseudo" transcript was constructed by combining the most severe consequence of each variant from the canonical and non-canonical transcripts within a gene[15]. For each method and scenario, we evaluated type I error and power at significance levels α = 0.001, 0.01 and 0.05. As an additional set of simulations, we repeated scenario 1 with an outcome prevalence of 2%.*

Manuscript Figure 1

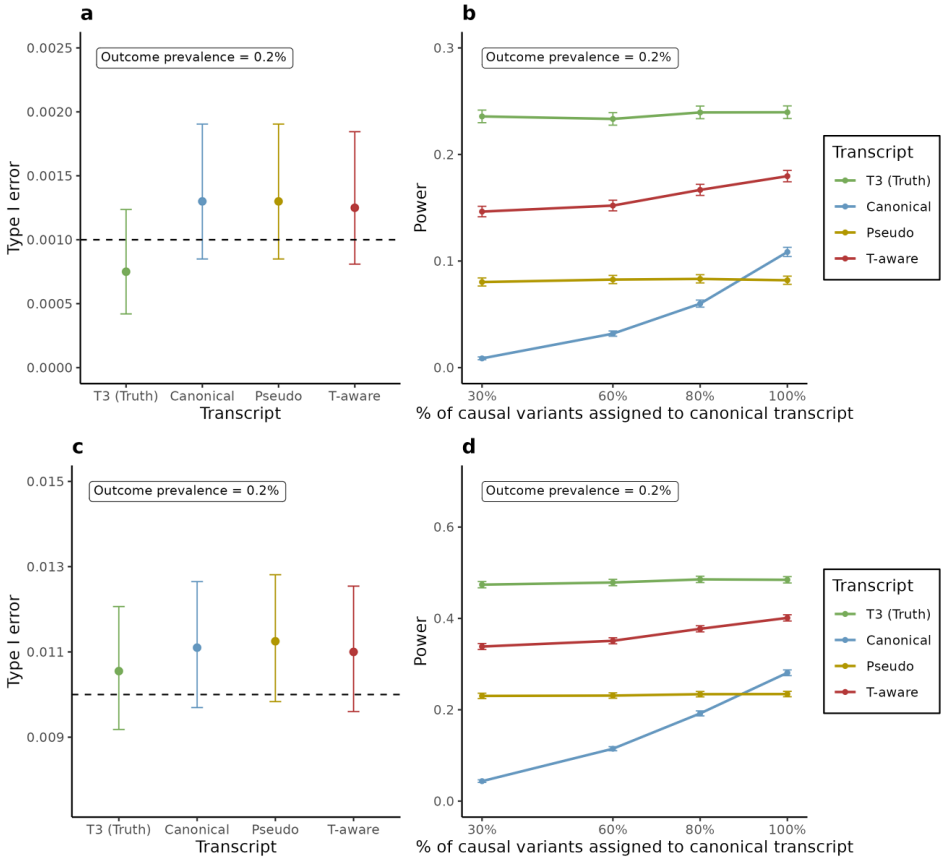
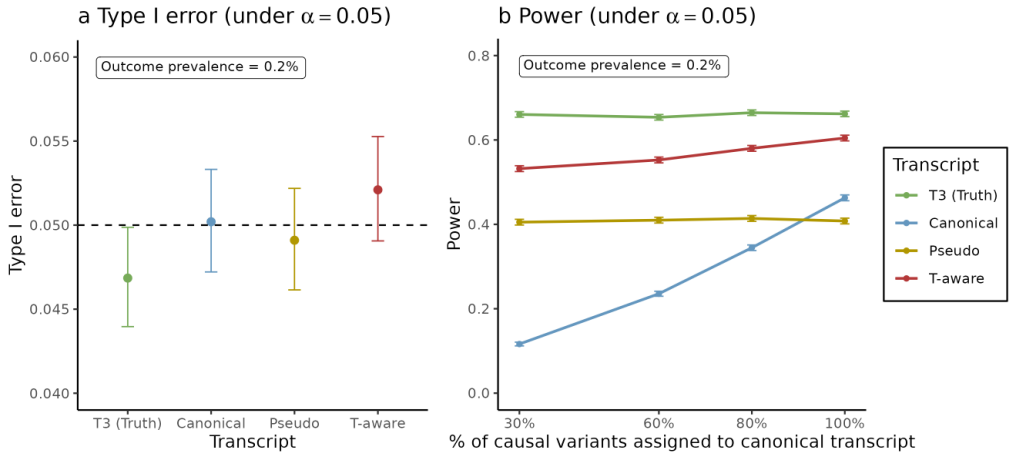


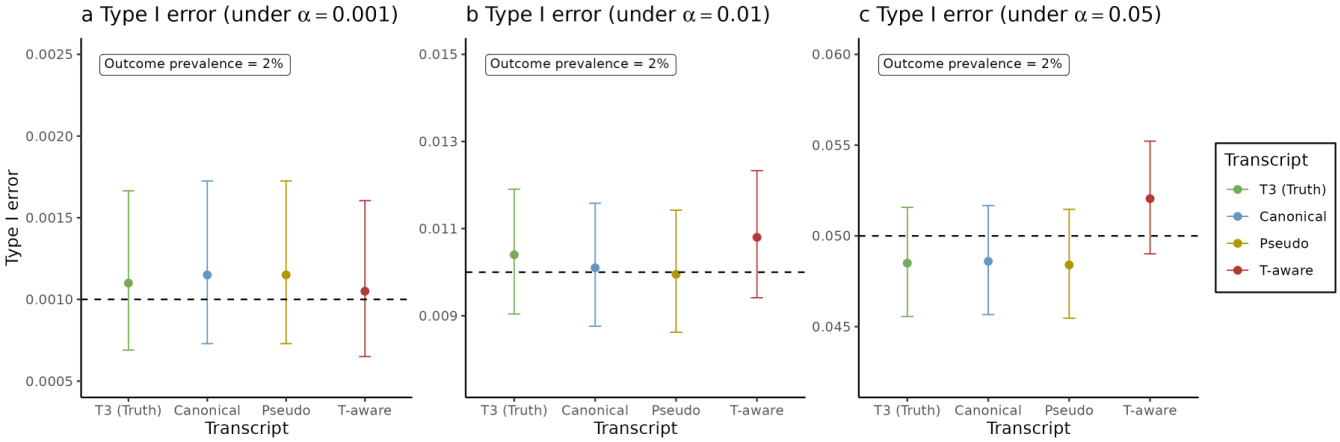
Fig.2 Type I error and power under different simulation scenarios.
***a** Estimated type I error rates for each method at α = 0.001, with 95% confidence intervals (CIs). **b** Power of each method as the percentage of causal variants assigned to the canonical transcript increases (30%, 60%, 80%, and*

100%) at α = 0.001, with 95% CIs. ***c*** *Estimated type I error rates for each method at α = 0.01, with 95% CIs.* ***d*** *Power at α = 0.01. T-aware is the proposed approach. T3 (Truth) represents an ideal case in which the most informative transcript is known in advance, and the test uses only T3. Canonical represents the test using only the canonical transcript. Pseudo represents the test using only the pseudo transcript.*

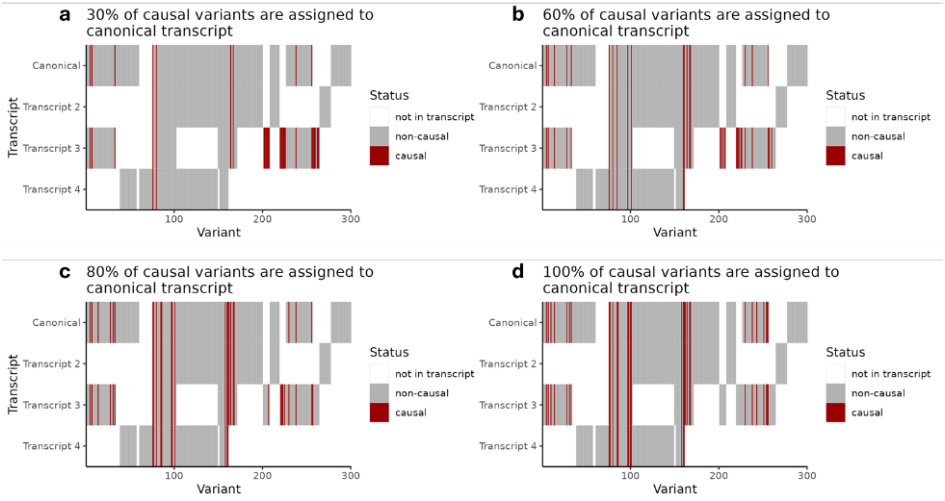
Supplemental Figures



Supplementary Figure 1: Type I error at α = 0.05 under a prevalence of 0.2%
a *Estimated type I error rates for each method at α = 0.05, with 95% confidence intervals (CIs).* ***b*** *Power as the percentage of causal variants assigned to the canonical transcript increases (30%, 60%, 80%, and 100%) at α = 0.05, with 95% CIs.*



Supplementary Figure 2: Type I error rates under a prevalence of 2%
a *Estimated type I error rates for each method at α = 0.001, with 95% CIs.* ***b*** *Estimated type I error rates for each method at α = 0.01, with 95% CIs.* ***c*** *Estimated type I error rates for each method at α = 0.05, with 95% CIs.*



Supplementary Figure 11: Simulated transcripts under different scenarios
a *Only 30% of the causal variants (colored in red) are assigned to the canonical transcript; transcript 3 contains all the causal variants. Non-causal variants are colored in gray.* ***b*** *About 60% of the causal variants are assigned to the canonical transcript.* ***c*** *About 80% of the causal variants are assigned to the canonical transcript.* ***d*** *The canonical transcript contains all causal variants.*

3. Line 78: "The proposed framework demonstrated strong reliability, revealing 46 associations previously reported in rare variant studies." Since the pLoF+pDM burden test has already been applied in WES studies and is included as one of the candidate masks in the current framework, replicating known associations does not necessarily validate the reliability of the proposed method.

Response: Thank you for your comment. We intended to show that the transcript-aware framework reproduces established rare-variant associations, confirming that the overall pipeline is correctly implemented and statistically well calibrated. We have revised the sentence in the manuscript.

Manuscript, Page 6, Line 106: *The proposed framework successfully reproduced 46 previously reported rare-variant associations, confirming proper calibration and consistency with prior studies.*

4. Line 180 paragraph: It is unclear what the authors intend to convey in this paragraph. Are they suggesting that transcript-aware masking improves cross-ancestry consistency? Or are they simply noting that the effects of pLoF+pDM variants are consistent across ancestry groups? If it is the latter, it is not evident how this observation is related to the proposed transcript-aware masking strategy.

Response: We agree that our original wording was unclear. We did not intend to suggest that the transcript-aware framework improves cross-ancestry consistency. Our goal was to show that the observed associations were consistent across ancestry groups, serving as a robustness check to confirm that the observed associations were not driven by population structure. We have revised the paragraph to clarify this purpose.

Manuscript, Page 12, Lines 223-226: *To assess the robustness of our results, we performed ancestry-stratified analyses (EUR-like vs. non-EUR-like), which showed consistent effect estimates and overlapping 95% CIs (**Supplementary Fig.9**), indicating that the observed associations were not driven by population structure.*

Minor comments:

1. Figure 1a: x-axis title: need to remove SNP?

Response: We thank the reviewer for carefully reviewing the manuscript. We have updated the figure with the correct x-axis label.

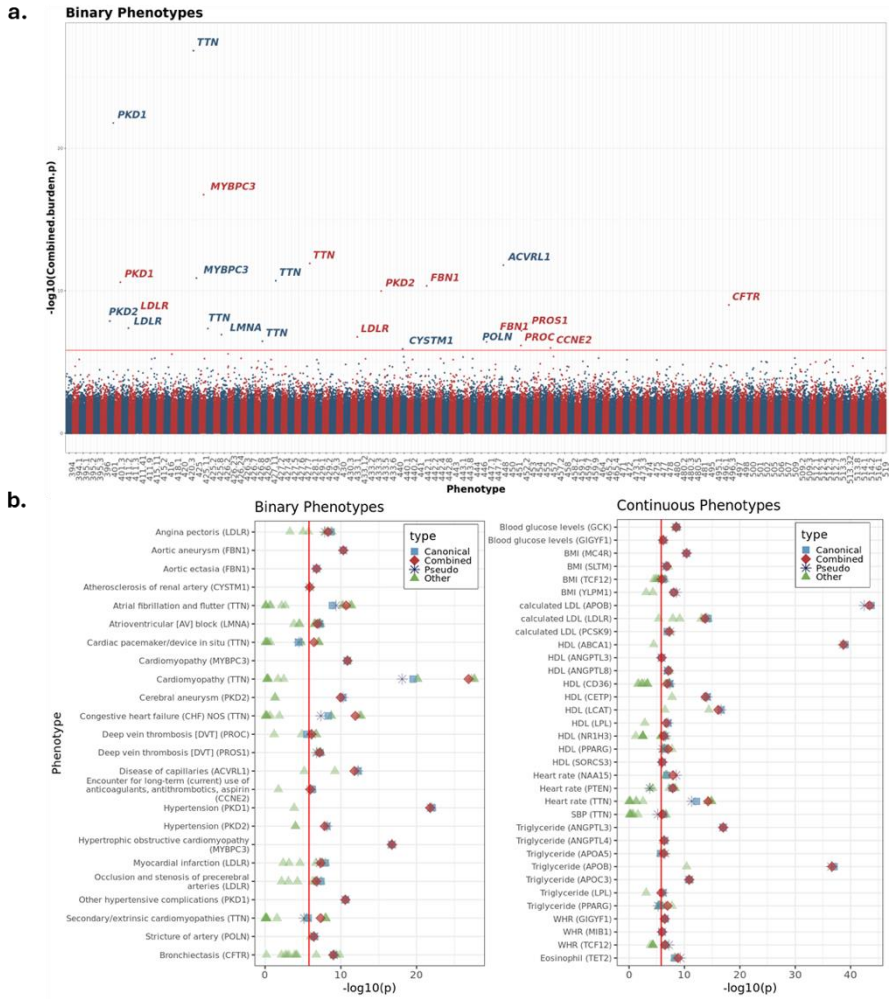


Fig.3 Significant associations identified under the transcript-aware testing framework.
a *Manhattan plot of exome-wide rare variant analyses for 118 binary phenotypes. The y-axis is the –log_{10} of the combined p-value ($P_{combined}$) for each test. For clarity, names of binary phenotypes are replaced by the corresponding phecode on the x-axis.* ***b*** *Significant associations by transcript type for binary and continuous traits. The x-axis displays the -log_{10}($P_{combined}$) for the corresponding associations. The red line represents a genome-wide p-value threshold ($P=1.5\times10^{-6}$) under a Benjamini–Hochberg FDR of 5% across all tests. To better illustrate the differences in results, transcript types are presented in different shapes and colors: red diamond – proposed method; blue square – canonical transcript; black star – pseudo transcript; green triangle – other transcript types. Abbreviations: BMI - body mass index; WHR - waist-to-hip ratio; SBP - systolic blood pressure; HDL - high-density lipoprotein cholesterol in serum or plasma; LDL - low-density lipoprotein cholesterol in serum or plasma.*

Reviewer #1 (Remarks on code availability):

The code lacks both usage instructions and example datasets, making it difficult to reproduce or evaluate the results.

Author Response: Thank you for the comment. We have updated our GitHub repository (https://github.com/seuchoi/meta_transcript). The repository includes detailed instructions and example simulated datasets for testing runs.

Reviewer #2 (Remarks to the Author):

The study from Zhang and colleagues proposes a transcript-aware burden test approach on 242,881 participants from the All of Us research program to assess the association with 129 cardiopulmonary traits. Genetic variants of interest included in the burden score are loss of function (LoF) and predicted deleterious missense (pDM). A key aspect of the work is the focus on transcripts, distinguishing canonical from non-canonical: the idea is simple, but important. Authors additionally created a "pseudo transcript", which combines the most predicted deleterious variants observed in the canonical and non-canonical transcripts for a certain gene. Methods are robust, data are well presented and treated according to rigorous quality control criteria. Overall, the study introduces a promising approach to better understand known associations and for discovering new ones in large heterogeneous cohorts.

Responses: We thank the reviewer for their recognition of the significance of our work. We appreciate the thoughtful evaluation of the study

I have a couple of comments for the Authors:

1. The logic behind the pseudo transcript creation is interesting, but I would add a cautionary note about the interpretation of the findings. In figure 1b, pseudo transcripts associations typically coincide with the combined analysis, sometimes association is slightly improved, but sometimes is also diluted, like in TTN and cardiomyopathy. This may be due to the incorporation of variants less pathogenic in practice. Indeed, despite the high confidence of the predictions, the actual impact of a genetic variant on a phenotype can be assessed only with in vitro experiments on appropriate models. This is important also for rare LoF, which sometimes are neutral, like the example of LRRK2 (Whiffin 2020 Nat Med). This also implies that variants, which are predicted to be less deleterious, could be pathogenic in fact. I would clarify this potential limitation.

   Response: Thank you for this valuable suggestion. We recognize the potential impact of including variants with uncertain or weaker effects. We have added this important potential limitation to the revised manuscript

   Manuscript, Page 12, Line 233-240: *We recognize some limitations of this study. For the proposed framework, although incorporating the pseudo-transcript increases sensitivity by integrating variants from canonical and non-canonical transcripts, it may also include variants whose functional impact is uncertain or weaker than predicted. This could introduce noise and dilute the association signal in some settings, particularly when the pseudo-transcript aggregates variants across transcripts that are not equally relevant to the phenotype. Predicted deleteriousness annotations are useful for interpretation but are not always accurate, and experimental validation remains essential to confirm the biological relevance of these variants*[46].

2. Finally, because the study is of general interest, to facilitate the less experienced reader in understanding the rich pipeline and supplementary materials, I would move the extended data figure 1 - Study Flow Chart to the main text.

Response: Thank you for the suggestion. We moved the Extended data figure 1 to the main text as Figure 1 in the revised manuscript.

Reviewer #3 (Remarks to the Author):

In this study, the authors present a transcript-aware rare variant association test. The research question is relevant, particularly in the context of aggregation tests. However, I have significant concerns about the All of Us analyses, and some sections of the manuscript would benefit from greater clarity.

Response: We appreciate the reviewer's evaluation of this manuscript. All comments and concerns related to the *All of Us* analyses have been addressed, and manuscript has been revised for clarity.

1. My main comment is about the genetic ancestry included in this study. If I understood correctly, the main analyses include all individuals, regardless of ancestry, except for the adjustment on common variants-based PCs. This choice is surprising given that population stratification is more pronounced for rare variants than for common variants, for which ancestry-aware meta-analyses are often performed. Even for the "ancestry-stratified" analyses, I have concerns about the justification of a "non-EUR" group.

   Response: Thank you for the comment. We analyzed all individuals together to maximize statistical power. Because most rare variants are extremely infrequent within any given ancestry group, splitting the sample into multiple ancestry-specific analyses would substantially reduce the number of carriers per variant, reducing the power of burden testing. This pooled analysis approach with PCs adjustment has been widely used in large-scale rare variant studies, including those using the UK Biobank and All of Us data (e.g., Backman et al., Nature 2021; Zhou et al., Nat Genet 2024).

   To empirically verify potential ancestry heterogeneity, we performed ancestry-stratified analyses (European-like vs. non-European-like groups) and observed consistent effect directions and magnitudes across groups as shown in **Supplementary** Figure 9, suggesting limited impact of residual stratification on our findings. Regarding the definition of the non-EUR group, the main reason for combining ancestries was the small sample size of several ancestry groups. For example, the South Asian-like and Middle Eastern-like groups included only about 2,300 and 500 participants, respectively (**Supplementary** Table 5). All of Us data use policies prohibit reporting association results with fewer than 20 carriers, which prevents us from performing ancestry-specific analyses.

2. Related to the previous point, why are there only 7 associations reported in Supp Fig 7 while 12 novel genetic associations are found? Is it because the other ones are not detected in EUR and non-EUR analyses?

   Response: As mentioned above, due to the data disclosure requirements of the All of Us Research Program, the presentation of summary statistics for subgroups with fewer than 20 individuals is prohibited. Therefore, all associations containing fewer than 20 individuals were removed from the supplementary figure. However, all 12 novel genetic associations were detected in both the EUR-like and non-EUR-like analyses.

3. The second main comment is about the additive value of looking at transcript-aware rare variant tests. Even if the benefit is highlighted in some examples, the overall gain is not clear. Looking at Figure 1, it seems for example that the signals identified in the combined analyses would also be detected in the proxy or canonical analysis. While the authors define 12 new signals, it is unclear whether this is due to the analysis of the All of Us dataset or to the transcript-aware analyses. It would be useful for example to give the percentage of genes detected only using the combined approach and not the canonical or pseudo transcripts.

Response: Thank you for the suggestion. The proposed framework utilizes both canonical and pseudo transcripts when available, so the percentage is not separable. However, we provided the information of the top transcript (transcript with the most significant p-value that is driving the association test result) in Table 1.

To further address the comment, we provide two summary tables below comparing the three approaches:

Across all associations		
	N significant tests	Total number of tests
T-aware (proposed)	58	1818862
Canonical-only	52	1739173
Pseudo-only	55	1818862

Among the 58 significant associations		
	Count	Percentage
T-aware < Pseudo	27	46.55%
T-aware < Canonical	23	39.66%
T-aware is minimum	18	31.03%

Using BH adjustment (FDR < 0.05) within each approach, T-aware identified 58 significant gene–phenotype associations, compared with 52 for canonical-only and 55 for pseudo-transcript. Three additional associations were detected only by T-aware (2 for *TTN* and 1 for *PPARG*). Notably, under the same minimum MAC threshold, the canonical-only analysis has fewer testable gene-phenotype pairs (1,739,173) than T-aware and pseudo-transcript-only (1,818,862). Among the 58 T-aware significant associations, T-aware produced smaller p-values than pseudo-transcript in 27 cases (46.55%) and smaller p-values than canonical-only in 23 cases (39.66%), and it was the most significant among all three approaches in 18 cases (31.03%). These results highlight the benefit of integrating evidence across transcripts. However, as mentioned before, the percentages are not separable and should be interpreted as identifying the subset where T-aware provides the strongest evidence beyond transcript-specific analyses, rather than a ranking of methods, because the proposed framework leverages both canonical and pseudo transcript information.

We would like to also highlight that this transcript-aware approach is particularly useful when a specific transcript is causally associated with a disease. As described in Manuscript Lines 105-135, the N2B isoform of *TTN* gene is highly expressed in the adult heart tissue compared to other transcripts. Using our framework, we observed that the associations between *TTN* and cardiomyopathy (Combined P-value[$P_{combined}$]=1.4×10^{-27}), atrial fibrillation and flutter ($P_{combined}$=2.0×10^{-11}), and congestive heart failure (CHF) NOS ($P_{combined}$=1.2×10^{-12}) showed substantially stronger than the *TTN* canonical transcript (**Figure 3b**). Conversely, when causal variants are distributed across transcripts, our approach achieves power comparable to conventional gene-based approaches.

We provided some of the discussions above in the revised manuscript:

Results and Discussion section, Lines 185-187: *Among the significant associations, the transcript-aware approach produced smaller p-values than the pseudo-transcript and canonical-only analyses in 46.55% and 39.66% of cases, respectively, and was the most significant across all three approaches in 31.03% of cases.*

Additionally, we stated in **line 215** that "*no instances were observed where the p-value from the canonical transcript was significant while the combined p-value was not*".

4. Related to the previous point, the authors compare the p-values between the different approaches. I am not sure what the authors mean by "a 100% gain in significance". Comparing the effect sizes of the tests might be more relevant.

Response: Thank you for pointing this out. We realize that the sentence is confusing. We compared the p-value from the proposed framework (1.4×10^{-27}) to the p-value from the canonical transcript (2.9×10^{-20}) to demonstrate the substantial decrease in significance after applying the framework. The purpose was to show the huge impact of considering more outcome-relevant (non-canonical) transcripts. We have revised the sentence.

Manuscript, Page 10, Line 192-194: *This substantial improvement of the statistical evidence demonstrated the value of non-canonical transcripts that may be more relevant to the outcome.*

In addition, I have additional comments:

5. From what I understand, whole-exome sequencing data and not WGS data are analyzed. Please change accordingly or clarify.

Response: Thank you for the comment. We analyzed the variants located in protein-coding (exonic) region extracted from the WGS data. We revised the corresponding sentence to make it clearer.

Manuscript, Page 4, Line 70: *We then applied the proposed framework to perform rare variant analyses using variants in exonic regions from the WGS data of over 240,000 All of Us participants.*

Manuscript Page 18, Line 311: *In this study, we utilized participants' age at enrollment, biological sex, predicted ancestries, and the first 16 principal components (PCs) computed by All of Us Research program as well as variants in protein-coding (exonic) regions derived from the WGS data (Supplementary Methods).*

6. In Table 1, it would be useful to indicate how many transcripts are present in the combined test.

Response: Thank you for your suggestion, we added the number of transcripts involved in the analysis in Table 1 and Supplementary Tables 1-2.

Trait	Phecode/ Concept id	$P_{combined}$ ($N_{transcript}$)	Gene	Top transcript				
				Transcript ID	Effect/OR (95%CI)	Carrier/ Cases	Carrier/ Controls	$P_{transcript}$
Binary phenotypes								

Atherosclerosis of renal artery	440.1	1.2E-06 (2)	*CYSTM1*	ENST00000644078	35.87 (13.05,92.08)	<20/347	23/217,569	1.2E-06
Encounter for long-term (current) use of anticoagulants, antithrombotics, aspirin	457	1.0E-06 (4)	*CCNE2*	C-ENST00000520509	6.36 (3.34,11.5)	<20/12100	58/186,177	5.5E-07
Occlusion and stenosis of precerebral arteries	433.1	1.7E-07 (9)	*LDLR*	C-ENST00000558518	4.78 (2.88,7.49)	<20/3625	353/223,246	4.0E-08
Stricture of artery	447.1	3.9E-07 (4)	*POLN*	C-ENST00000511885	5.43 (3.24,8.57)	<20/501	720/217,569	3.0E-07
				Transcript ID	**Effect/OR (95%CI)**	**Carrier/All samples**		***P*$_{transcript}$**
Quantitative phenotypes								
BMI	903124	8.3E-09 (4)	*YLPM1*	Pseudo	0.52(0.35,0.69)	97/237,177		3.0E-09
		1.2E-06 (11)	*TCF12*	ENST00000559609	0.52(0.32,0.72)	93/237,177		3.1E-07
WHR	-	3.2E-07 (11)	*TCF12*	Pseudo	0.47(0.3,0.64)	98/218,607		6.1E-08
		1.1E-06 (2)	*MIB1*	C-ENST00000261537	0.15(0.09,0.21)	760/218,607		1.1E-06
Heart rate	903126	1.1E-08 (5)	*NAA15*	Pseudo	0.69(0.46,0.92)	69/237,809		3.0E-09
		1.3E-08 (8)	*PTEN*	ENST00000700021	0.85(0.57,1.14)	45/237,809		4.0E-09
Triglycerides	3022192	1.1E-07 (18)	*PPARG*	ENST00000652522	1.1(0.72,1.48)	26/99,299		1.8E-08
HDL	3007070	1.0E-06 (2)	*SORCS3*	C-ENST00000369701	-0.76(-1.06,-0.45)	28/96,908		1.0E-06

Novel associations with an FDR Q-value < 0.05 ($P=1.5\times10^{-6}$) among 58 identified gene-trait associations. Values for $P_{combined}$ were obtained by combining test results from all transcripts based on the proposed framework, and $N_{transcript}$ is the number of transcripts.

7. Some supplementary materials (ex Supp Fig 9) are not referenced in the text.

Response: Thank you for the comment. We added a summary sentence of Supplementary Fig 9 (currently Supplementary Fig 10) in the main text, with more descriptions in the Supplementary Results section.

Manuscript, Page 9, Lines 172-176: *For significant gene–phenotype associations, penetrance among rare variant carriers ranged from 1.2% to 35% (Supplementary Fig.10). The highest penetrance was observed for PROC variants associated with deep vein thrombosis (35%), followed by PKD1 (32%) and PKD2 (23%) for hypertension, whereas most other associations showed low-to-moderate penetrance (< 10%).*

Supplementary Results: *The strongest association was observed between "disease of capillaries" which includes hereditary hemorrhagic telangiectasia (HHT) and the ACVRL1 gene, whose mutations are a known cause of HHT. This association had an OR of 52.41 (transcript ID: ENST00000550683, 95% confidence interval (CI) [24.05, 102.90], $P=5.3\times10^{-13}$), with mutation carriers showing a phenotypic penetrance of 10% (Supplementary Fig.10).*

8. Can the authors provide a list of all the phenotypes tested for association along with the same size? It is not so clear how

are the clusters used, and how the phenotypes were selected.

Response: Yes. We provide the sample sizes and number of cases for each phenotype in **Supplementary Table 8**. Details of the steps are described in the "Methods: Study population and phenotypes" and "Supplementary Methods" sections. Our approach is similar to that of *Sun et al., Nature (2022).* In brief, we conducted hierarchical clustering to identify highly correlated phenotypes. Their corresponding cluster memberships are provided in **Supplementary Table 7**. The number of clusters was determined by the highest average silhouette width (**Supplementary Fig.12**). We then selected one phenotype from each cluster in consultation with clinicians to make sure the selected phenotype is a good representative for the cluster.

Supplementary Table 8. *Summary statistics for the 118 binary phenotypes we selected from hierarchical clustering.*

Group	Trait	Case, n(%)	Number of Observations
Circulatory	Rheumatic disease of the heart valves	2987(1.34)	223128
	Mitral valve stenosis and aortic valve stenosis	502(0.23)	220180
	Nonrheumatic mitral valve disorders	5426(2.41)	225104
	Nonrheumatic aortic valve disorders	3996(1.79)	223674
	Nonrheumatic tricuspid valve disorders	1071(0.49)	220749
	Abnormal heart sounds	4416(1.95)	226606
	Hypertension	13364(5.65)	236364
	Other hypertensive complications	2624(1.16)	225388
	Myocardial infarction	6305(2.89)	218004
	Angina pectoris	5669(2.61)	217368
	Aneurysm and dissection of heart	245(0.12)	211944
	Other acute and subacute forms of ischemic heart disease	474(0.22)	212173
	Pulmonary embolism and infarction, acute	3368(1.46)	230535
	Chronic pulmonary heart disease	2041(0.89)	229208
	Cardiomegaly	4000(1.7)	235600
	Precordial pain	2305(1.23)	186679
	Carditis	2543(1.08)	234505
	Endocarditis	854(0.37)	232814
	Cardiomyopathy	4758(2.01)	236724
	Hypertrophic obstructive cardiomyopathy	260(0.11)	232220
	Secondary/extrinsic cardiomyopathies	549(0.24)	232509
	Other cardiomyopathy	245(0.11)	232205
	Atrioventricular [AV] block	2164(1.18)	184007
	Second degree AV block	368(0.2)	182211
	Atrioventricular block, complete	693(0.38)	182536
	Bundle branch block	3323(1.79)	185166
	Abnormal electrocardiogram [ECG] [EKG]	7701(4.06)	189544
	Other cardiac conduction disorders	495(0.27)	182338
	Cardiac pacemaker/device in situ	3005(1.63)	184848
	Paroxysmal supraventricular tachycardia	3209(1.73)	185052
	Atrial fibrillation and flutter	10033(5.23)	191876
	Cardiac arrest and ventricular fibrillation	728(0.4)	182571
	Arrhythmia (cardiac) NOS	5541(2.96)	187384
	Premature beats	3747(2.02)	185590
	Tachycardia NOS	7524(3.97)	189367
	Congestive heart failure (CHF) NOS	6530(2.93)	222751
	Heart transplant/surgery	512(0.24)	216863
	Abnormal function study of cardiovascular system	2480(1.13)	218701
	Symptoms involving cardiovascular system	4756(2.15)	220977
	Intracranial hemorrhage	1239(0.55)	224540
	Subdural hemorrhage	386(0.17)	223632
	Occlusion of cerebral arteries, with cerebral infarction	387(0.17)	223633
	Cerebral atherosclerosis	301(0.13)	223547
	Occlusion of cerebral arteries	4192(1.84)	227438
	Cerebral ischemia	6595(2.87)	229841
	Cerebral aneurysm	1069(0.48)	224315
	Acute, but ill-defined cerebrovascular disease	546(0.24)	223792
	Atherosclerosis	3693(1.67)	221715
	Atherosclerosis of renal artery	347(0.16)	217916
	Atherosclerosis of the extremities	1499(0.68)	219068
	Vascular insufficiency of intestine	501(0.23)	218242
	Aortic aneurysm	2353(1.07)	219922
	Aneurysm of iliac artery	225(0.1)	217794
	Arterial dissection	243(0.11)	217812
	Aneurysm of other specified artery	425(0.19)	217994
	Peripheral vascular disease	6875(3.06)	224444
	Raynaud's syndrome	1676(0.76)	219245
	Other specified peripheral vascular diseases	202(0.09)	217771

	Arterial embolism and thrombosis	603(0.28)	218341
	Polyarteritis nodosa and allied conditions	969(0.44)	218538
	Stricture of artery	501(0.23)	218070
	Aortic ectasia	1573(0.72)	219142
	Disease of capillaries	409(0.19)	218677
	Noninfectious disorders of lymphatic channels	2162(0.89)	242881
	Phlebitis and thrombophlebitis	1218(0.65)	187696
	Deep vein thrombosis [DVT]	4266(2.24)	190443
	Chronic venous hypertension	355(0.19)	186578
	Varicose veins	5684(2.96)	192330
	Hemorrhoids	11526(5.51)	209282
	Encounter for long-term (current) use of anticoagulants, antithrombotics, aspirin	12100(6.1)	198277
	Encounter for long-term (current) use of antiplatelets/antithrombotics	1840(0.98)	188017
	Hypotension	6461(2.85)	226498
	Iatrogenic hypotension	405(0.18)	220174
	Hemorrhage NOS	437(0.2)	220206
	Blood vessel replaced	812(0.37)	220581
	Circulatory disease NEC	3684(1.65)	223453
Respiratory	Acute sinusitis	10730(5.49)	195281
	Acute pharyngitis	13701(7.07)	193926
	Acute laryngitis and tracheitis	591(0.33)	180816
	Nasal polyps	931(0.52)	178747
	Chronic pharyngitis and nasopharyngitis	5139(2.8)	183735
	Chronic laryngitis	401(0.23)	178147
	Paralysis/spasm of vocal cords or larynx	850(0.48)	178596
	Acute and chronic tonsillitis	2235(1.24)	180060
	Chronic sinusitis	10992(5.73)	191960
	Epistaxis or throat hemorrhage	2335(1.29)	181118
	Throat pain	1090(0.61)	179773
	Pneumonia	10662(4.82)	221365
	Viral pneumonia	470(0.23)	206504
	Pneumonia due to fungus (mycoses)	354(0.17)	206388
	Bronchopneumonia and lung abscess	220(0.11)	206254
	Influenza	2186(1.03)	211846
	Asthma	24689(11.1)	222328
	Chronic obstructive asthma	892(0.46)	193303
	Emphysema	3205(1.64)	195616
	Bronchiectasis	1392(0.72)	193803
	Bronchitis	5634(2.77)	203049
	Acute bronchospasm	534(0.28)	193433
	Lung disease due to external agents	433(0.2)	217475
	Pneumonitis due to inhalation of food or vomitus	871(0.4)	218031
	Postinflammatory pulmonary fibrosis	1481(0.68)	219098
	Other pulmonary inflamation or edema	577(0.27)	217593
	Empyema and pneumothorax	1298(0.59)	218478
	Pleurisy; pleural effusion	4870(2.18)	223806
	Respiratory failure, insufficiency, arrest	6207(2.78)	223086
	Respiratory insufficiency	262(0.12)	217134
	Pulmonary insufficiency or respiratory failure following trauma and surgery	390(0.18)	217262
	Wheezing	2685(1.62)	165417
	Painful respiration	3985(2.39)	166717
	Abnormal chest sounds	364(0.22)	163096
	Shortness of breath	27779(14.58)	190511
	Hypoventilation	1349(0.57)	236875
	Orthopnea	270(0.11)	235796
	Disorders of diaphragm	478(0.2)	236004
	Abnormal results of function study of pulmonary system	580(0.27)	218410
	Solitary pulmonary nodule	6596(2.91)	226365
	Hemoptysis	1078(0.45)	240914
	Other diseases of respiratory system, not elsewhere classified	3599(1.53)	234867

Selected phenotypes were grouped by their general category, either circulatory or respiratory. In total, we had 76 phenotypes from circulatory and 42 from respiratory. The corresponding cases (%) and sample size were also included in this table.

9. Were the analyses in UKB performed in a similar way as in All of Us?

Response: The UKB results we obtained are publicly available on Genebass (https://app.genebass.org/). In the main manuscript, we are looking at the possible replication of the association in an independent cohort. According to Karczewski et al., GeneBass provides rare variant associations across 4529 phenotypes available in UK biobank, using multiple masking strategies and statistical tests. For our comparison, we focused on loss-of-function variants (most severe consequences) from burden test, which is conceptually similar to our pseudo-transcript approach, although transcript-

specific information was not incorporated in their analyses. Therefore, while the analytic pipelines are not identical, the GeneBass results allow us to assess whether the same gene shows evidence of association in an independent dataset. We agree that fully harmonized analyses in UKB would provide a stronger replication framework. Therefore, we acknowledge in the Discussion that further replication and validation studies will be necessary to confirm our findings.

Manuscript, Page 12-13, lines 241-245: *In addition, while some of our findings were supported by comparisons with existing literature and UKB summary statistics[40], there remains a pressing need for large-scale, ancestrally diverse cohorts and harmonized analytic approaches to enable comprehensive, comparable, and generalizable genetic analyses in the future.*

REVIEWER COMMENTS

We thank the editor and all reviewers for their time to review the revision.

Reviewer #1 (Remarks to the Author):

The authors have addressed my previous comments and conducted additional simulation studies to evaluate the performance of the proposed method. The simulations generally demonstrate that the proposed approach adequately controls type I error rates. However, the power comparisons suggest that the T-aware approach is substantially more powerful than the two alternative methods (canonical and pseudo).

1. I am concerned that these results are somewhat misleading, as the simulation settings appear to favor scenarios in which the proposed method is expected to perform optimally. In the All of Us data analysis, however, the canonical and pseudo approaches often yielded smaller p-values than the proposed approach. According to the authors' response to Reviewer 2 (Comment 3), the T-aware method achieved the minimum p-value in only 31% of significant associations, and the total number of significant associations was not substantially different across methods.

   To provide a more balanced and comprehensive evaluation, the authors should conduct power simulations under a broader range of scenarios. This would help clarify the conditions under which the proposed method achieves superior power, as well as situations in which its performance is comparable (e.g., via CCT) or potentially less favorable relative to existing approaches. Such analyses would offer a more realistic and informative assessment of the method's practical utility.

   We thank the reviewer for evaluating and identifying the additional components to improve our previous version of the simulation. To fully illustrate the performance of the proposed approach under different settings, we updated our simulations and conducted additional simulations under different causal configurations (full details attached in the end).

   We believe that the true causal architectures underlying the associations investigated in the *All of Us* data likely contain all causal configurations (described below) considered in our simulation studies, and possibly additional patterns as well. When the canonical transcript contains most of the causal variants, it is expected to be the most outcome-relevant transcript, so the canonical-only approach tends to have the smallest p-value (Simulation causal configuration 3). When causal variants are distributed across multiple transcripts, the pseudo-only approach is expected to have the largest power and therefore the smallest p-value (Simulation causal configuration 4). When causal variants are concentrated in a non-canonical transcript, the proposed method performs the best (Simulation causal configuration 2, our previous simulation results). When a gene has only one transcript, all methods produce the same p-value.

   Collectively, these observations suggest that the associations observed in *All of Us* reflect a mixture of causal configurations, which helps explain why the proposed method achieved the smallest p-value in only 31.03% (=18/58, including 1 tie with canonical and pseudo) and 29.31% (=17/58, excluding 1 tie) of the 58 significant associations. For your reference, we provide the count of each approach having minimum p-value among 58 significant associations below:

Canonical	Combined	Pseudo	Tie	Total
17	17	8	16	58

As you can see, none of the methods consistently has the minimum p-value, potentially reflecting a mixture of underlying causal configurations. However, as shown in both the simulation results and **Fig. 4**, the proposed method remains competitive even when the underlying configuration is not the most favorable to it, because the framework also leverages information from the canonical-only and pseudo-only approaches. This means that for very significant associations, we would expect the proposed method to also detect those detected by canonical-only and pseudo-only approaches. Additionally, one limitation of our study is that we restricted the analysis to cardiopulmonary traits (mentioned in our **Limitation** section line 281-283). Expanding the analysis to a broader range of traits may allow us to observe more causal configurations and thus find other strong cases to show the benefit of the framework. We believe that all the reasons above explain why the total number of significant associations was not substantially different across methods.

In reality, because we do not know the true causal configuration, the proposed framework can serve as a practical tool for prioritizing transcript-level signals while maintaining competitive overall performance. Notably, **Fig. 4** also shows that differences in p-values across transcripts are most apparent when the proposed method yields the smallest p-value (e.g., *TTN* and *PPARG*). By contrast, when the canonical-only or pseudo-only approach produces a more significant p-value than the proposed method, the differences in p-values are usually less apparent (e.g., *LDLR* and *LMNA*).

In addition to the newly added simulations, we added **supplementary Fig. 3** to further illustrate the behavior of the p-values produced by the Cauchy Combination test (CCT) framework and the robustness of the proposed framework to variation in the number of non-associated transcripts. The result suggests that the proposed framework is reasonably robust when at least one transcript carries a very strong signal.

The updated Method, Results and Supplementary sections are included below:

Methods

Simulation

*We conducted simulations with n=200,000 individuals for a hypothetical gene with 300 rare and ultra-rare variants under the general rare variant analysis settings (**Supplementary Methods**). Of these variants, 20% were rare (minor allele frequency [MAF] ≈ 0.005%–0.05%) and 80% were ultra-rare (MAFs ≈ 0.0005%–0.005%). For simplicity, we assumed no linkage disequilibrium among variants, which is reasonable for rare and ultra-rare variants. Variants were mapped to four transcripts (T1–T4), where T1 was the canonical (longest) transcript. We considered four general causal configurations to evaluate the proposed framework, including one null setting for type I error and three settings for power.*

In configuration 1 (scenario 1), none of the transcripts contained causal variants, and this setting was used to evaluate type I error. In configuration 2 (scenarios 2-5), the total number of causal variants was fixed to 30, with all causal variants increasing disease risk (all in the same effect direction). T3 was treated as the most outcome-relevant transcript by containing all causal variants.

*Power was assessed by varying the proportion of causal variants assigned to the canonical transcript while T3 always contained all causal variants. These proportions were 30%, 60%, 80%, and 100%, corresponding to 9, 18, 24, and 30 causal variants in the canonical transcript (**Supplementary Fig. 12**). In configuration 3 (scenarios 6-8), the canonical transcript was the most outcome-relevant transcript that always contained all causal variants, and T3 was treated as noise with no causal variants. In configuration 4 (scenarios 9-11), causal variants were randomly located across transcripts. Configurations 3 and 4 were evaluated under varying numbers of causal variants (10, 20 and 30; **Supplementary Fig. 13**).*

*For each of the eleven scenarios, we repeated the simulations 20,000 times. The outcome was binary with a prevalence of 0.2%, roughly corresponding to a disease that occurs about 1 in 500 individuals, such as hypertrophic cardiomyopathy[53] (**Supplementary Methods**). We compared four approaches (1) canonical-only, (2) pseudo-only, (3) transcript 3 (T3)-only, and (4) our proposed transcript-aware method, all using the burden test. Here canonical-only represented a standard approach that used a single transcript per gene without pre-selection. The "pseudo" transcript was constructed by combining the most severe consequence of each variant from the canonical and non-canonical transcripts within a gene[15]. In configuration 2, T3-only represented an idealized case in which the most informative transcript was known in advance. For each method and scenario, we evaluated type I error and power at significance levels $\alpha = 0.001$, 0.01 and 0.05. As an additional set of simulations, we repeated scenario 1 (the null setting) with an outcome prevalence of 2%. Furthermore, to better understand the behavior of the combined p-value, we plotted the final aggregated p-value from the proposed framework using hypothetical p-value vectors outside the simulation setting (details in **Supplementary Methods**).*

Results and Discussion

Simulation Study

*Simulation results comparing four approaches (canonical-only, pseudo-only, transcript 3 [T3]-only, and the proposed transcript-aware [T-aware] method) across various causal configurations were summarized in **Fig.2** and **Supplementary Fig.1**. At $\alpha = 0.001$, all methods showed well-controlled type I error. The T-aware method kept the false positive rate close to the nominal level, while the T3-only approach was slightly lower than nominal (**Fig.2 a**). At $\alpha = 0.01$ (**Supplementary Fig.1 a**), type I error rates were close to nominal for all methods, with overlapping confidence intervals (CIs). Similar patterns were observed at $\alpha = 0.05$ (**Supplementary Fig.1 b**), with the T3-only approach being slightly more conservative and T-aware being slightly above the nominal level. To further assess these behaviors, we performed additional simulations under a higher outcome prevalence of 2% (**Supplementary Fig.2**). Consistent with the results under 0.2% prevalence, all methods controlled type I error well at $\alpha = 0.001$ and $\alpha = 0.01$, with T3-only still slightly conservative and T-aware slightly above nominal at $\alpha = 0.05$. The pattern of the p-values at $\alpha = 0.05$ may reflect both Monte Carlo variation from a finite number of replicates and case-control imbalance present in rare variant analyses.*

*When causal variants were concentrated in the non-canonical transcript T3 (causal configuration 2), the T-aware method was more powerful than the canonical-only and pseudo-only approaches across all scenarios, with non-overlapping CIs for the estimated power (**Fig.2 b** and **Supplementary Fig.1 c-d**). The difference in power was larger when the canonical transcript contained fewer causal variants. Notably, even when the canonical transcript contained all causal variants, the*

canonical-only approach was still less powerful because it included many additional non-causal variants. As expected, the highest power was achieved when the most outcome-relevant transcript (T3) was known in advance and analyzed alone.

*When all causal variants were concentrated in the canonical transcript and T3 served as a noise transcript (causal configuration 3), the canonical-only approach outperformed T-aware (**Fig.2 c** and **Supplementary Fig.1 e-f**), as the canonical transcript was the most informative one under this configuration. However, T-aware remained competitive and achieved the second highest power as this configuration can be treated as a special case of configuration 2, where a canonical transcript was the most outcome-relevant transcript instead. In causal configuration 4, where causal variants were randomly distributed across transcripts, the benefit of leveraging multiple transcripts became more apparent. As expected, the pseudo-only approach showed the highest power, while T-aware had the second highest power, and the difference between them depended on the total number of causal variants (**Fig.2 d** and **Supplementary Fig.1 g-h**). Overall, these results showed that power depended on multiple factors, including the causal configuration under our simulation design, while T-aware remained competitive with the ideal single-transcript approach across settings. More importantly, because the truly outcome-relevant transcript and the distribution of causal variants are generally unknown in reality, T-aware may serve as a practical surrogate for the ideal approach.*

*Outside the simulation setting, we examined the behavior of the final combined p-value from the proposed framework under an extreme scenario where only one transcript was significant, and all others were non-significant (**Supplementary Methods**). We found that adding more non-significant transcripts gradually reduced the combined signal. However, the magnitude of this decrease depended heavily on the smallest p-value in the set, as well as the magnitudes of the remaining p-values (**Supplementary Fig.3** and **Supplementary Results**). With an extremely low p-value, the combined result remained significant even after many additional transcripts were included, suggesting that the framework is reasonably robust to dilution by non-significant transcripts. Although this extreme scenario is unlikely to occur in practice, it provides useful insight into the behavior of the proposed framework.*

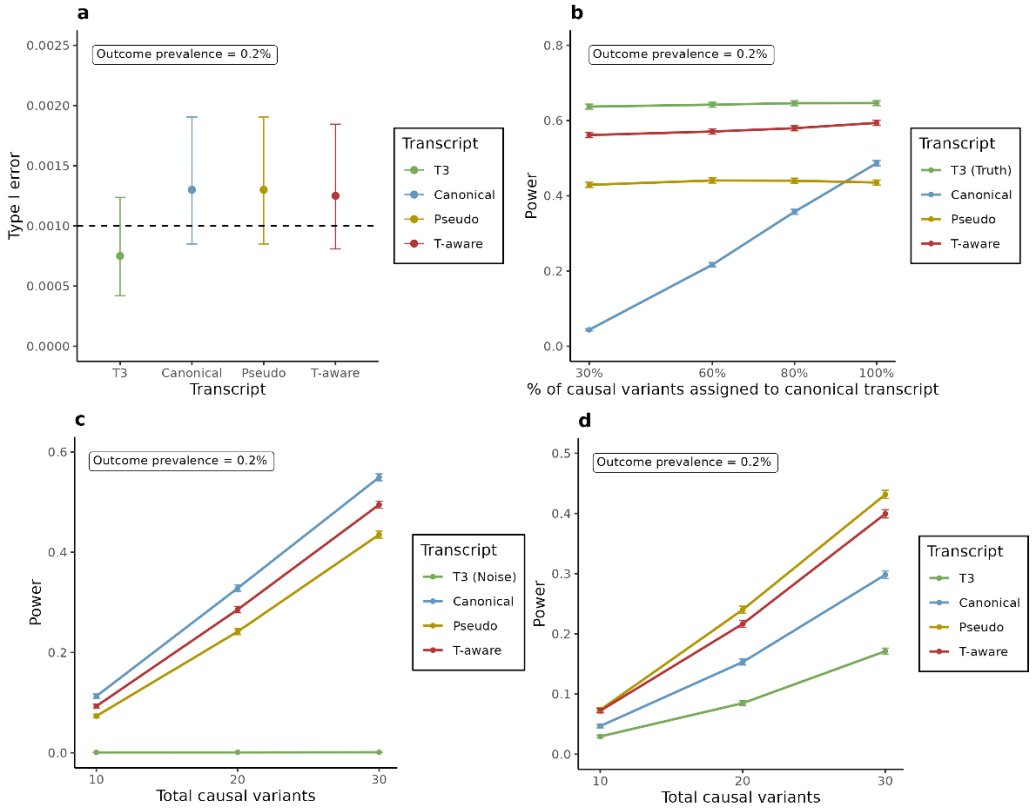


Fig.2 Type I error and power under different simulation settings.

a Estimated type I error rates for each method at α = 0.001 in causal configuration 1, with 95% confidence intervals (CIs). b Power of each method in causal configuration 2 as the percentage of causal variants assigned to the canonical transcript increases (30%, 60%, 80%, and 100%) at α = 0.001, with 95% CIs. Transcript 3 is the most outcome-relevant transcript. c Power of each method in configuration 3 as the total number of causal variants increases (10, 20, and 30) at α = 0.001, with 95% CIs. Transcript 3 does not contain causal variants. d Power of each method in configuration 4 as the total number of causal variants increases at α = 0.001, with 95% CIs. T3 represents the test using only transcript 3. Canonical represents the test using only the canonical transcript. Pseudo represents the test using only the pseudo transcript. T-aware is the proposed approach.

Supplementary - Methods

6. Simulation

6.1 Genetic data

We simulated gene-level rare-variant data for n=200,000 individuals and 300 variants using a four-transcript gene structure consisting of one canonical transcript (T1) and three alternative transcripts (T2 to T4). T2 to T4 were created as partial subsets of T1 with additional transcript-specific exons, so that variants could be shared across transcripts or unique to a specific transcript. Minor allele frequencies (MAFs) were generated to mimic rare variant architecture, with 80% sampled from an ultra-rare range and 20% from a rare range. Genotypes were then simulated independently under a binomial model (assuming no linkage disequilibrium). We considered four causal configurations with multiple scenarios:

- *For causal configuration 1 (scenario 1), no causal variants were assigned.*

- *For causal configuration 2 (scenarios 2 to 5), a total of 30 causal variants were selected from transcript T3, with a prespecified proportion also shared with T1 (30%, 60%, 80%, and 100%). This makes T3 the most outcome-relevant transcript.*
- *For causal configuration 3 (scenarios 6 to 8), all causal variants were selected from T1 while excluding variants in T3, making T1 the most outcome-relevant transcript and T3 a non-causal transcript (noise). This setting can be viewed as a special case of the configuration 2 design in which a single transcript is most strongly related to the outcome, namely T3 in S2-S5 and the canonical transcript in S6-S8. The expected results can be inferred by switching T1 and T3 in the results for configuration 2. Therefore, to avoid providing redundant simulation results, we reduced the length of T1, and T3 was further modified to be outcome-irrelevant in S6-S8.*
- *For causal configuration 4 (scenarios 9 to 11), causal variants were sampled randomly from all simulated variants across the gene.*

In all scenarios, all causal variants were set to have the same effect direction. Larger weights were assigned to rarer variants using weights based on the Beta (0.5, 25) density. For causal configurations 3 and 4 (S6-S11), the transcript architecture was modified relative to configurations 2 (S2-S5): the alternative transcripts (T2 to T4) were defined as smaller subsets of the canonical transcript and included more transcript-specific exons. The canonical transcript (T1) was slightly narrowed. This design created a more distinct transcript structure with reduced overlap across transcripts, allowing us to evaluate the proposed framework under a broader range of causal configurations. Overall, the simulation was designed as a general rare variant analysis for illustration purposes.

6.2 Binary outcome

For each simulated genetic dataset, we generated a binary outcome from a logistic regression based on the individual's genetic risk score. Specifically, for each individual i, the genetic score was calculated as the weighted sum of variant dosages $\sum_j G_{ij}\beta_j$ for all j variants, where the weights were given by the simulated effects described in the previous section. This score was then converted to a probability p, where $p = \frac{1}{1+e^{-\eta}}$ and η was the sum of the intercept and genetic score. We chose the intercept to ensure the prevalence matched the prespecified values (0.2% and 2% in our case). Finally, the binary phenotype was sampled from a Bernoulli distribution using p.

6.3 Evaluating the behavior of the aggregated p-value

To further assess the behavior of the aggregated p-value, we plotted the final aggregated p-value from the proposed method using hypothetical p-value vectors outside the context of the simulation study. We plotted -log$_{10}$(p) under an extreme scenario for illustration purposes. Specifically, we considered a vector of p-values in which only one p-value is significant, while the remaining p-values are similar in magnitude and all non-significant at the 0.05 level. In this setting, each p-value represents the test result for a specific transcript. We have:

$$p = (\underbrace{p_{min}}_{\text{The only } p-\text{value that is significant.}}, \underbrace{p_2, p_3, p_4 \dots \dots p_k}_{\text{Rest of } p-\text{values that are all non}-\text{significant.}})$$

$$\text{where } p_{min} \in \{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 0.01, 0.05\}$$

$$p_2, p_3, p_4 \dots \dots p_k \text{ will all be a value around } \{0.9, 0.6, 0.06\}$$

Two example p-value vectors are:

- *A set of 3 p-values (3 transcripts), where the minimum p-value is 10^{-6} and the other two are around 0.9: p = (10^{-6}, 0.9010175, 0.8935225).*
- *A set of 5 p-values, where the minimum p-value is 10^{-6} and the remaining four are around 0.6: p = (10^{-6}, 0.6046895, 0.6001637, 0.6010480, 0.5961736).*

We then plotted the final aggregated p-value from the proposed method against the total number of p-values to examine how many non-significant p-values (transcripts) would need to be included before the overall significance is heavily affected.

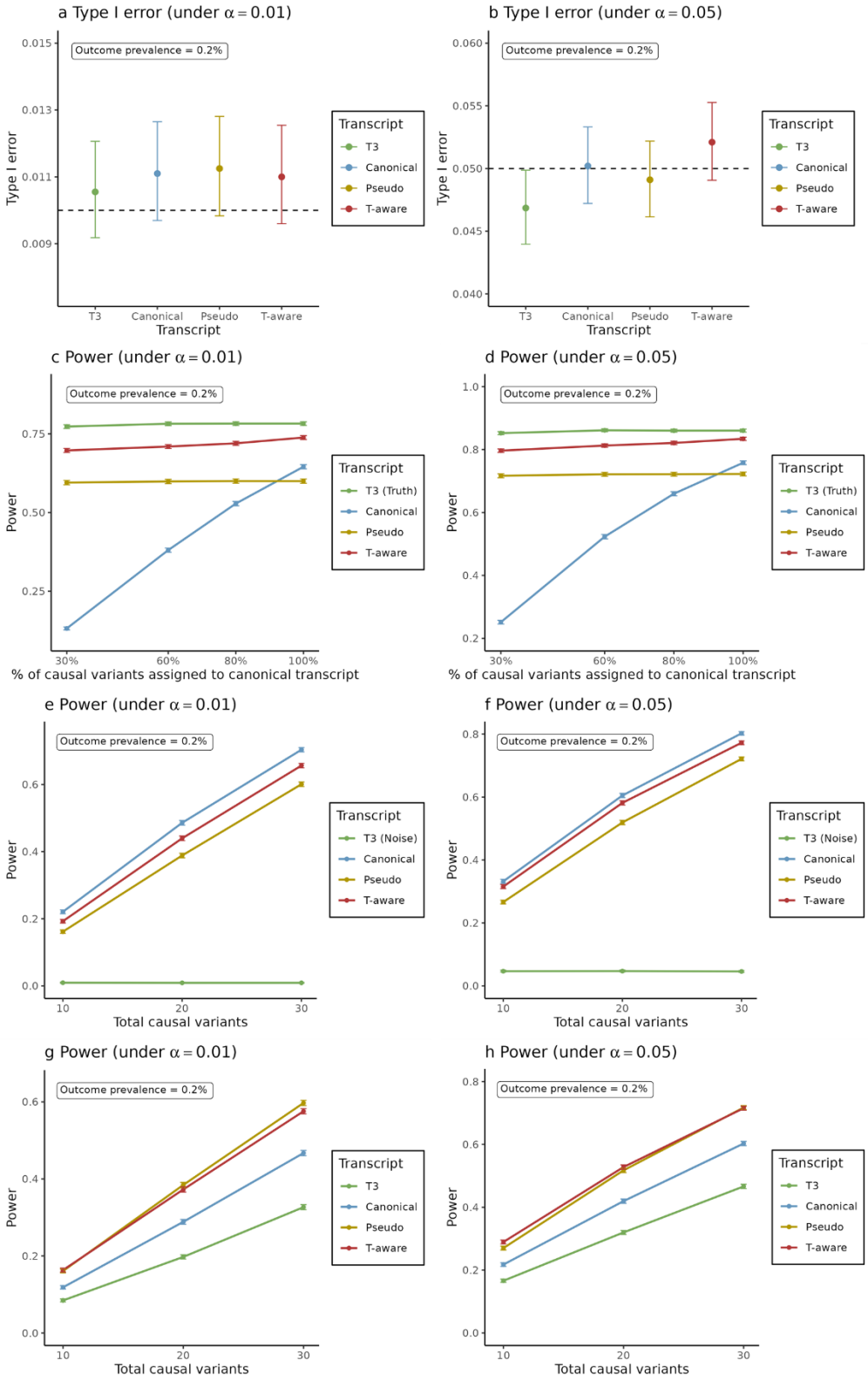
Supplementary - Results

2. Simulation

2.1 Evaluating the behavior of the aggregated p-value

*We generated **Supplementary Fig.3** to show how the final aggregated p-value changed as additional non-significant transcript-level p-values were included in the set. Overall, the combined significance weakened gradually as the number of non-significant transcripts increased, but the change in magnitude depended strongly on both the minimum p-value in the set and the magnitude of the remaining p-values. When one transcript had a very small p-value, such as 10^{-6}, 10^{-5}, 10^{-4}, or 0.001, the aggregated p-value ($P_{combined}$) stayed well below 0.05 even after including up to 20 transcripts. This suggested that the proposed method was relatively robust to the addition of many non-significant transcripts in these settings. In contrast, when the minimum p-value was more modest (e.g., 0.01 or 0.05), the final significance became much more sensitive to the remaining p-values. This loss of significance was most apparent when the other p-values were large, approximately 0.6 or 0.9. However, this represented an extreme setting in which only one transcript was informative, and all others were non-significant. In practice, more than one transcript may have carried useful signal.*

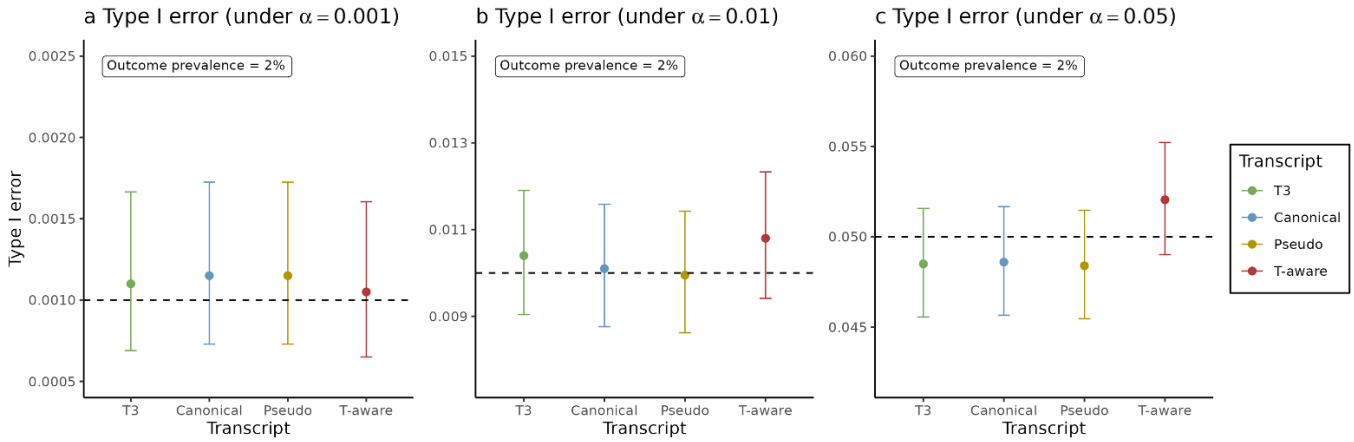
Section 4. Supplemental Figures



Supplementary Figure 1: Type I error and power under a prevalence of 0.2%

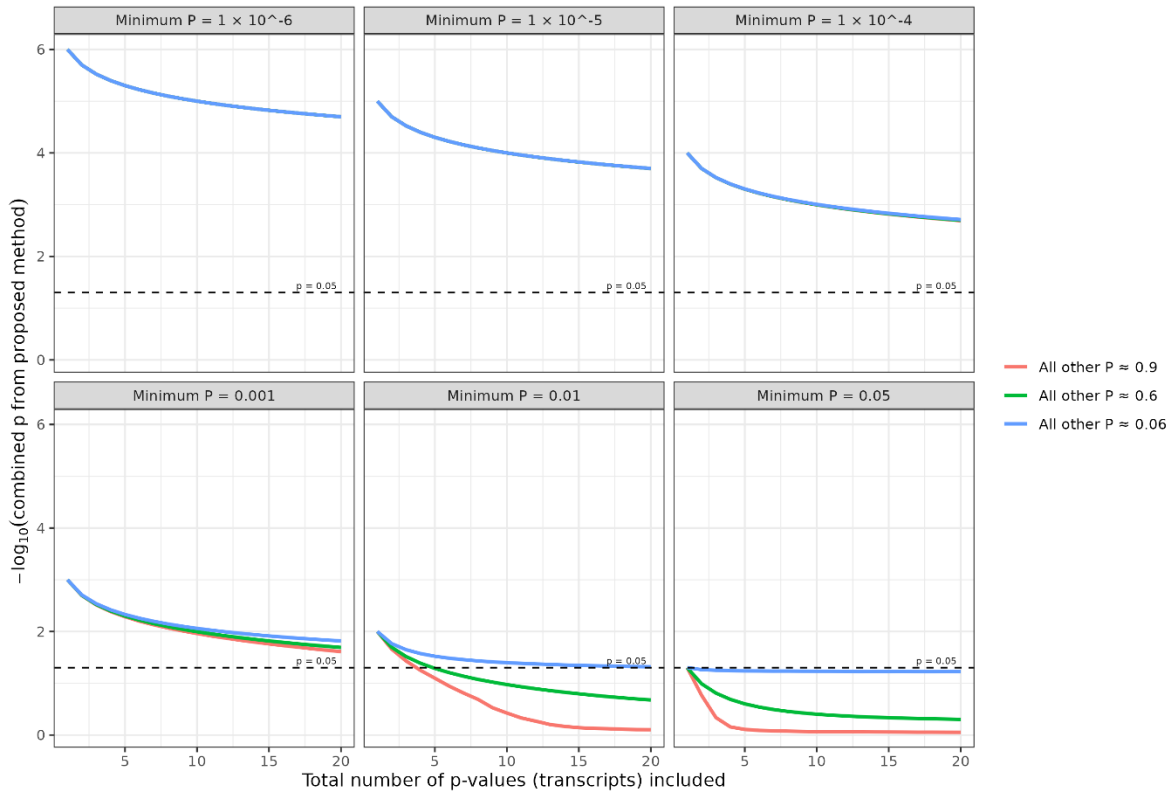
a *Estimated type I error rates for each method at α = 0.01, with 95% confidence intervals (CIs).* ***b*** *Same as panel a, but at α = 0.05.* ***c*** *Power as the percentage of causal variants assigned to the canonical transcript increases (30%, 60%, 80%, and 100%) at α = 0.01, with 95% CIs. Transcript 3 is the most outcome-relevant transcript.* ***d*** *Same as panel c, but at α = 0.05.* ***e*** *Power as the total number of causal variants increases (10, 20, and 30) at α = 0.01, with 95% CIs. The canonical transcript always includes*

all causal variants. Transcript 3 does not contain causal variants. ***f*** *Same as panel e, but at α = 0.05.* ***g*** *Power as the total number of causal variants increases (10, 20, and 30) at α = 0.01, with 95% CIs. Causal variants are randomly located across transcripts.* ***h*** *Same as panel g, but at α = 0.05.*



Supplementary Figure 2: Type I error rates under a prevalence of 2%

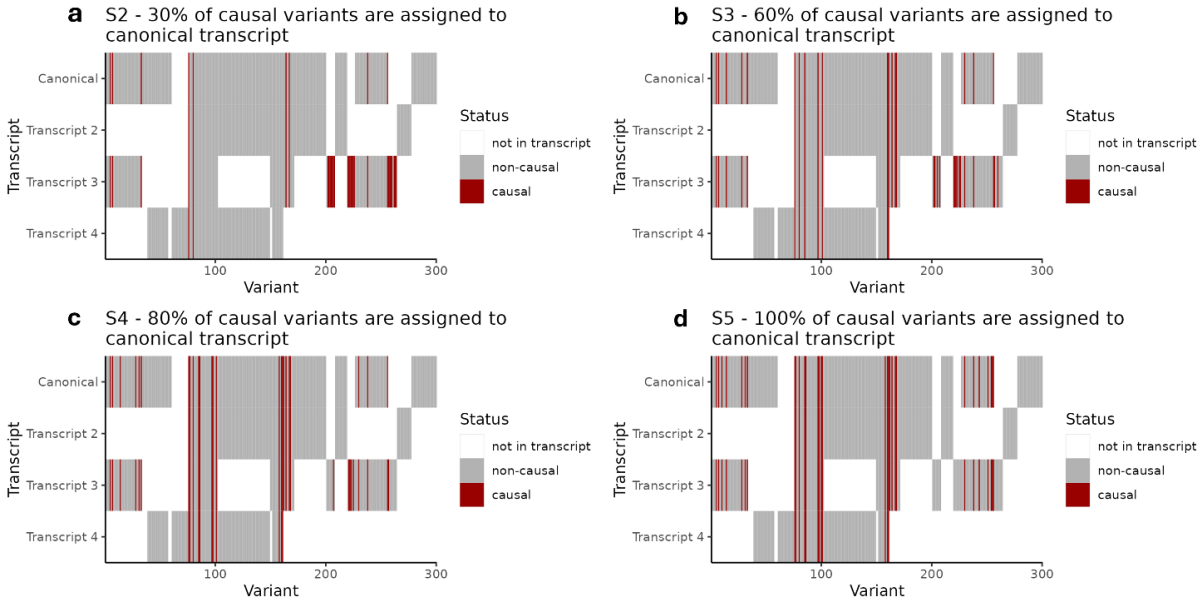
a *Estimated type I error rates for each method at α = 0.001, with 95% CIs.* ***b*** *Estimated type I error rates for each method at α = 0.01, with 95% CIs.* ***c*** *Estimated type I error rates for each method at α = 0.05, with 95% CIs.*



Supplementary Figure 3: Behavior of the final combined p-value from the proposed method versus the number of transcripts included

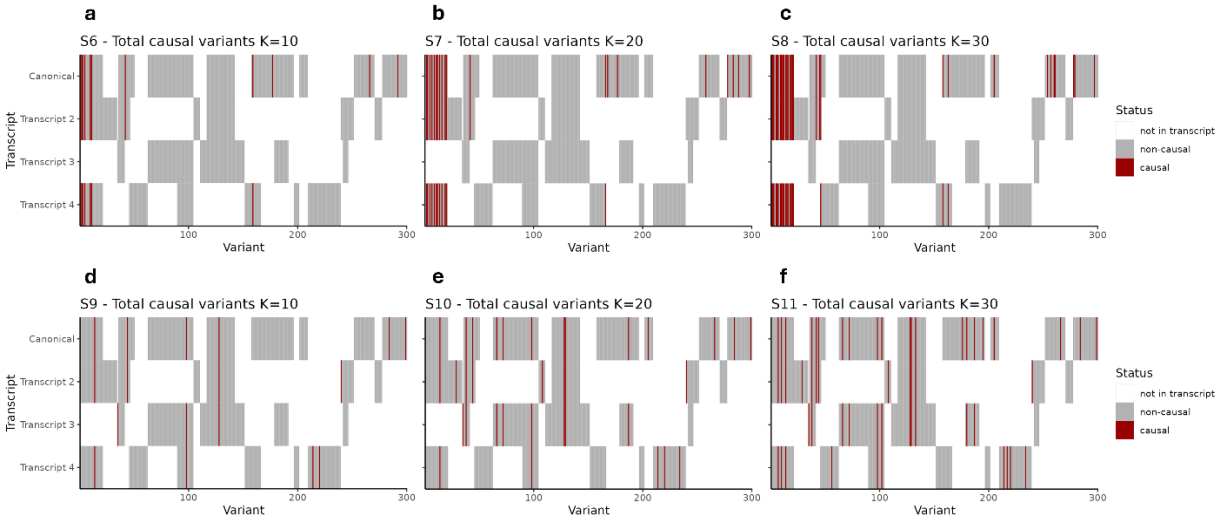
Each panel shows $-\log_{10}$ of the final combined p-value from the proposed method as a function of the total number of transcript-level p-values included in the set. For illustration, hypothetical p-value vectors were considered outside the simulation setting under an extreme scenario in which only one

transcript-level p-value was informative, and all remaining p-values were non-significant (at the a = 0.05 level). Panels correspond to different minimum p-values in the set (10^{-6}, 10^{-5}, 10^{-4}, 0.001, 0.01, and 0.05). Colored lines indicate settings in which all remaining p-values were approximately 0.9, 0.6, or 0.06. The horizontal dashed line marks the nominal significance threshold of p = 0.05.



Supplementary Figure 12: Simulated transcripts under different overlapping rates in causal variant configuration 2

***a** Only 30% of the causal variants (colored red) are assigned to the canonical transcript; transcript 3 contains all causal variants. Non-causal variants are colored gray. **b** About 60% of the causal variants are assigned to the canonical transcript. **c** About 80% of the causal variants are assigned to the canonical transcript. **d** The canonical transcript contains all causal variants. S2-S5 denote scenario numbers. Note: the figure shows examples from one random seed.*



Supplementary Figure 13: Simulated transcripts under different causal variant configurations

***a** All 10 causal variants (colored red) are assigned to the canonical transcript (causal configuration 3); transcript 3 contains no causal variants. Non-causal variants are colored gray. **b** Same as panel a, but*

with 20 causal variants in total. ***c*** *Same as panel a, but with 30 causal variants in total.* ***d*** *A total of 10 causal variants are randomly distributed across transcripts (causal configuration 4).* ***e*** *A total of 20 causal variants are randomly distributed across transcripts.* ***f*** *A total of 30 causal variants are randomly distributed across transcripts. S6–S11 denote the scenario numbers. Note: the figure shows examples from one random seed.*

2. Reviewer #1 (Remarks on code availability):

   The quality of the software package needs improvement, as it currently depends on a file path that is not accessible to external users. Specifically, the code includes a reference such as:
   source('/UKBB_200KWES_CVD/GENESIS_adaptation_source.R')

   The authors should revise the implementation to ensure that all required scripts and dependencies are properly included within the package structure, making it fully self-contained and reproducible for external users.

   We thank the reviewer for helping us improve the quality and professionalism of the code package. We have updated the corresponding GitHub site so that it no longer uses an external source.