nature portfolio

Peer Review File

Population-Specific Polygenic Risk Scores Developed for the Han Chinese

Corresponding Author: Professor Pui-Yan Kwok

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.

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

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)
The manuscript by Chen et al presents results of GWAS, heritability, genetic correlations, colocalization and PRS over a new biobank data of ~500k individuals within Taiwan (Taiwan Precision Medicine Initiative, TPMI) trying to identify population-specific genetic risk as a major source of risk in this population. The main addition to community is the resource of GWAS/PRS summary results over hundreds of ICD-based phenotypes that will likely provide another non-European centric genomic resource; I commend the authors for providing the summary results as open resource. As the community greatly needs more diverse biobank data, I am supportive towards the publication of this work. However, I find the majority of the claims being overstatements not fully supported by analyses presented and I encourage the authors to rephrase their results.

1. Population labels. Terms such as "East Asian (EAS)", "Han Chinese", "Asian" need to be defined rigorously clearly labeling which ones are referring to genetically inferred ancestry versus self-reporting. For genetically inferred ancestries, it is important to clarify exactly what computational method used and how many individuals were excluded and did not meet the criteria. For self-reporting, it is important to describe the selection criteria for the inclusion in the biobank. The manuscript erroneously uses these different concepts interchangeably.

2. TPMI Biobank representativity of Taiwan. The authors compare disease rates in the biobank versus the National Health Insurance Research Database in Taiwan(NHIRD) finding a correlation in the prevalences of r=0.64. I disagree with their conclusion that a correlation of 0.64 is proof that that TPMI represents well the population. Such a small correlation could be indicative of ascertainment issues and/or problematic phenotypic based on phecodes. Moreover, the case proportion of TPMI seems deflated across the board as compared to NHIRD prevalence (Figure 1a) thus suggesting that TPMI samples a healthier snapshot of the population. The authors could look at distributions of age/sex/locations/etc to describe the main drivers of ascertainment in TPMI. Even if TPMI is not a perfect representation of general population (which it seems to be the case), it is still a great resource and highly useful for discoveries. I suggest the authors describe this in the text of section discussing Figure 1.

3. A central theme of the work is that population-specific genetic risk is a major driver for complex traits. I find this claim not supported in the data. For example, cross-population correlations are very high for most traits (Table S6) (lines 508-514). Furthermore, the vast majority of the loci have been reported before (only 77 out of 1,139 are novel). Can the authors report which loci have evidence of heterogeneity across populations? How many of those heterogeneities can be ascribed to true differences in genetic architectures (as claimed on lines 513-514) and how many are due to LD-tagging or differences in frequencies?

4. GWAS sample size. Methods section (line 752) suggests that GWAS was performed after excluding 100k unrelated individuals; I found that problematic in unnecessarily restricting power. Moreover, line 764 notes that only 248k (out of 500k) subjects are "unrelated"; it is highly unusual to have so many related samples in the biobank. Can the authors clarify sample sizes for each analysis? Can the authors clarify excluded sample sizes based on various criteria? It seems like 20k subjects were excluded based on quality control alone (line 729).

5. Comparison with other Taiwanese biobank (TWB). The authors do a limited comparison with TWB when discussing external validation of PRS models. Since TWB has similar sample size as TPMI, it would greatly enhance the manuscript if the authors also compare the GWAS/h2/correlations across the two biobanks. Furthermore, if the manuscript presents PRS that would be the best for Taiwanese individuals, it seems like combining results across the biobanks would surely provide superior accuracy than just one biobank.

6. TPMI-derived PRS outperformed UKB derived PRS when applied to EAS ancestries; this is not surprising and great to see. But wouldn't that suggest that a joint model combining EUR_PRS with EAS_PRS would be the best?

7. The authors find that PRS for various traits correlate with number of visits and hospitalization duration explaining ~9% of the variation. Part of this could be due to non-genetic factors due to residual stratification in the PRS. How much are other factors like age, sex, hospital location, explaining in this trait? Were they adjusted for in the analyses of lines 572-578? This may impact the statement of lines 643-644 as non-genetic factors may drive predictability. Also important to note in text that number of visits and/or days in the hospital are a poor pricy for "disease burden". Also note that TPMI may not be a perfect proxy for all Taiwan (see comment above).

8. GWAS replication rates range from 19.81% to 74.4%. To what do the authors ascribe this? A replication rate of 10-30% seems extremely low. Can the authors stratify replication rates according to statistical power expected (as function of frequency and effect size in existing work)? I find the statements on lines 438-439 not assuring at all.

9. Lines 440-446 report that only 77 (out of 1,139) are novel associations with text noting how these highlight the uniqueness of these loci in TPMI (lines 457-459). Which of the 77 have support for heterogeneity (i.e. different causal effect) or are simply due different statistical power for discovery?

10. Heritability estimates seem too low and different from other reported results (lines 468-476). First, the authors should clarify they report SNP-heritability results (not narrow-sense results). Second, how do these results compare to other biobanks? For example h2-snp of 0.3/0.2 for height/bmi seems too low. Since TPMI is not the only large-biobank from Taiwan, the authors could compare heritability results across biobanks.

11. lines 487-490: unclear and too broad of a statement.

12. lines 501-506: how are these findings significant for clinical applications?

13. I commend the authors for describing PRS performances in the context of upper limits of SNP-heritabilities (lines 516-535).

14. line 593: how about other PGS from the catalog that include larger sample sizes over UKB? Is TPMI-derived PRS still better than the most accurate European PRS from PGSCatalog?

15. lines 673-680: this is too broad of a claim. There is no formal validation in the manuscript for this hypothesis.

16. A chip-GWAS was performed (lines 737-739). Any details on how the results of that scan were used to minimize bias? Was there a "sample collection site"-GWAS performed as well?

17. Distribution of relatedness in TPMI is needed.

18. Unclear how PRS was trained. Why was only 100k of TPMI used for PRS instead of the whole sample of 500k? or at least the 250k that are unrelated? (lines 839-841)

19. lines 863-883: it is unclear what are the sample sizes of the cohorts used in this study. How many participants from All of US where used for this study? Why was a PRS trained on 80k samples in TPMI compared with a sample size of 120k from TWB? How about a PRS that uses all samples together; wouldn't those be the best PRS for this population?

Referee #2

(Remarks to the Author)
The authors of the manuscript "Population-Specific Polygenic Risk Scores Developed for the Han Chinese" have done a tremendous amount of work in the Taiwan Precision Medicine Initiative. This is an important resource and I believe the manuscript will be a worthy contribution to the field, showing how diverse biobanks can bring both biological insights and necessary tool development for precision medicine. Overall, the paper would be greatly improved if the authors took up some more real estate going through the specific examples as to why novel discoveries were made (such as is it related to allele frequency differences or different epidemiology of that specific trait), as well as clarifying some vague statements.

Right now there is a lot of work that has been included, but almost everything is up to the reader to do a deep dive of every gene without some exemplars demonstrating how important this work is. I have separated out my comments below into major and minor concerns. Some things I have noted as not necessary to the paper, but would be a nice addition.

- The manuscript would be substantially strengthened by delving into the reasons why there are novel signals and relationships within the TPMI compared to much larger non-Han Chinese biobanks. This could take the form of allele frequency and/or effect size comparisons, or comparison of prevalence of different outcomes between the biobanks. While the included analyses are substantial and a very worthy contribution to the field, going a bit deeper and trying to explain it would be a huge contribution and make a better argument for the need for diverse biobanks. This would be necessary throughout the manuscript. Right now a lot of it is focused on discovery, which again while this would create a resource of known associations, if the central premise of the paper as per the abstract is that diverse biobanks are needed, more real estate should go into explaining why that is beyond the yield of novel associations. Overall, the paper would be greatly strengthened by putting their results into context of the field at large, such as pop-specific vs multi-pop PRS or what is known about the genetics of specific outcomes (such as HBV).
- Was the correlation between the National Health Insurance Research Database different by trait category? This may identify some bias in terms of the recruitment of the 16 medical centers. Also, were there differences by sites? (This last part is not necessary for the manuscript, I just think it would be an interesting piece tied to why TPMI is different or as a citation for future TPMI papers.)
- The manuscript would be strengthened in terms of readability if some of the technical details were integrated into the main text. For example, In line 434 the authors talk about replication rate with other EAS GWAS, but there is no mention of where these GWAS come from, nor is it in the Methods. It took me looking at Extended Data Fig. 1 to find that they are pulled from the GWAS Catalog. How were these selected/filtered from the overall database?
- Another example is the use of "independent" without saying what independent actually means (R2<X, etc). While some of these details are found in the Methods, there's no harm (and would actually help the reader) to have some of them in the main results text.
- The investigators use the threshold of 5X10E08 for genome-wide significance, however there are >700 GWAS being conducted and therefore this does not account for multiple comparisons across analyses. Would the changing of this threshold change the conclusions at all? If not, then I would just say that. If so, how?
- Some of the heritability estimates seem substantially lower than other biobanks/studies, such as height which has a heritability of 32.3%. What is the interpretation of this?
- I would love a bit more explanation of how to interpret/interesting findings from Figure 3 with an expansion of the results in the main text. Right now it's a bit of a data dump, which while understandable makes it hard to parse what readers should find super interesting.
- The use of phecodes that are highly correlated and even could be nested (such as psoriasis, psoriasis and related disorders, psoriasis vulgaris, psoriasis athropathy) is a bit misleading when it comes to the genetic correlation and clustering. With so much overlap in cases, of course they will genetically overlap. For these analyses I could see two possible avenues. One would be to remove highly correlated phecodes (such as T2D and 7 different T2D w/specific manifestations/symptoms) to get a better idea of how these clusters inform each other, OR to focus on what is different or specific to these phecodes. For example, are there differences in genes associated with T2D w/peripheral circulatory disorders or T2D with renal manifestations?
- The above concern re:redundant phecodes may be particularly concerning with the PRS training using multiple phecode-specific PRS. As I interpret the manuscript currently, the authors only compared the phecode-specific PGS developed in TPMI with external validation in the other large-scale Taiwan study and UKB/AoU. PRSmix has the possibility of overfitting which would be demonstrated with this external validation.
- For the PGS methods, please add the actual numbers for the AUC instead of relative improvements (see lines 541-542). Relative improvement is not as relavant to interpretation as the actual AUC.
- The language around external validation and comparison of PRS models is both vague and misleading. For example, lines 561-562 state that the "TPMI-derived PRS models also consistently outperformed UKB European-derived models when applied to EAS", which is shown in Figure 6. However, the confidence intervals for the TPMI-derived and UKB-derived almost always overlap for most of the traits shown in Figure 6. The only exceptions seem to be T2D (although only for the two Taiwan studies), gout (again only for Taiwan-based studies), and migraine (only TWB?). It is fine to say that they performed just as well, or seem to perform better *although the confidence intervals overlap*.
- Are the PGS adjusted for any other variables, such as age, sex, and study? If so, are the R2 values the incremental R2 of just the PGS or are they the overall performance of the model including these variables? These two have very different interpretations and I didn't see those details in the methods.
- The "Impact of Genetic Risks on Overall Disease Burden" is similarly vague and overstating the results through omission. For example, lines 579-580 state that the cardiometabolic cluster contributes to 1.14% of clinical visits and 3.22% of hospitalizations. What is omitted is that is only comparing the top 5% to the bottom 5% and that the values decreased substantially when looking at the top vs bottom 20%. This distinction is important as it isn't that the overall PRS accounts for that proportion looking at linear relationships from both factors, but only the extremes of the PGS. I don't think this weakens the authors points at all! It just needs a bit of explanation to put into context. Also, it would be interesting to know if these results differ by study site to identify selection bias or overall differences between the 16 medical centers.
- Were any sensitivity analyses conducted such as matching on age or using a longitudinal design on age to account for the 4th limitation in lines 663-665? That may address some of the issues with younger participants not having time to develop the outcome at all? Or any inclusion/exclusion criteria for age?
- The authors state in lines 673 that the study validates the common belief that population-specific risk-prediction models perform well. However, they did not compare their pop-specific PRS to a multi-pop PRS, so cannot make the claim that this means we need pop-specific methods. This is also somewhat counteracted with the paragraph directly after that shows that

sometimes the population doesn't matter (equal performance if UKB or TPMI-derived). These two paragraphs could be combined and put into better context with what is known in the field for the development of PGS for these outcomes.

Minor
- It would be helpful to add the correlation line to the plot in Figure 1A to make it consistent with the main text (r=0.64, line 426).
- The y-axis for Figure 1B is off (more space between 100-200 than 200-300). This data may be better presented in a table with sample sizes and then mean/median/range or case/control counts to get a sense of scale. This is especially important when considering all the splits of analyses for the PRS development/validation.
- The manuscript reads several time as if the investigators are surprised to have a signal for HBV. This is not surprising given the prevalence of this infection and the sequelae associated with chronic infection. If anything, this is an important contribution to the GWAS of infectious disease and speaks to the strength of TPMI re:different epidemiology for specific traits. This could be expanded on a bit more to add a bit of punch to the discussion/conclusions.
- Why were the specific cutoffs for PRS used in lines 522-524? For example, why look for models explaining 3% of the trait variance? Or 0.55 for the AUC?
- Please remove AUC from the plots showing r2 and heritability for PGS. They are not comparable and the plots are misleading, especially since the AUC is often way above the h2 (because they are not the same). AUC can be a separate plot.
- Is the use of "epidemic" on line 616 a typo? Did you perhaps mean "endemic"? (I would also expand in the discussion as to why geographically and ancestrally diverse biobanks are needed, as the epidemiology of HBV is very different in Taiwan versus other places.)
- Another source of diverse eQTL data could be MAGE (https://github.com/mccoy-lab/MAGE), just in case you hadn't seen that yet.

Referee #3

(Remarks to the Author)
Chen et al. present the Taiwan Precision Medicine Initiative (TPMI), which is a truly remarkable dataset. I enjoyed reading the manuscript, and seeing how far the genetics field has come. TPMI includes over half a million Taiwanese residents, making it possibly the largest east asian genetic dataset with electronic medical record data. I think this is a very impressive study, which addresses a critical gap in precision medicine for non-European populations, and will likely have a very significant impact on the human genetics field. The publicly available online resource and GWAS summary statistics will likely be widely used in future genetic studies. There are many interesting analyses in the paper, and generally I found them to be of high quality, e.g. GWAS and PGS analyses were impressive. It is also a very well written paper. However, I suggest reconsidering the emphasis on predicting disease burden in Taiwan, as I believe selection biases may significantly impact these numbers, possibly rendering them meaningless. Besides this comment, most of my other comments are relatively minor, and should be straightforward to address in a revised manuscript. The comments below are aimed at improving the manuscript.

1. I would appreciate at least some comments on selection biases, and how they might bias the analyses, and perhaps consider employing approaches to address these. These could include inverse probability weighting (see e.g. Schoeler et al. Nat Hum Behav 2023.

1.a). This study evaluated the “genetic impact on overall disease burden”. I believe this estimate could be very biased due to selection biases, including participation bias, as well as time-related biases (e.g. right and left censoring).

1.b) Similarly, heritabilities and PGS accuracies will likely be impacted by selection biases. Interestingly, you do adjust for case-control ascertainment. Perhaps you could comment on this?

2. I was impressed by the large number of PGS methods that you used. However, I would have liked to see a clearer comparison between the different PGS methods. You claim that LDpred2 outperformed other methods, but it’s hard to see from the supplementary tables.

3. What does “Researchers requesting access to the 909 individual genotyping and EMR data can do so on a collaborative basis.” mean? This statement seems rather ambiguous to me, and I think it could be improved.

4. In line 645 you say that “implementing genetic risk-based health management strategies may decrease the disease burden”. Although I don’t necessarily disagree, and I don’t know what the future holds, I would be surprised if the genetic predictors radically improved prediction of disease burden compared to standard risk factors (e.g. age, sex, smoking status, BMI, etc.)

5. In line 764-6 you note that you used “PLINK2 for GWAS among the unrelated subset (n = 248,754)51, and these statistics were used for heritability and genetic correlation estimation.” I hope you also used these for the PGS, as LDpred2 and most other methods assume the GWAS summary statistics are obtained using a logistic regression or linear regression, and not from mixed models (e.g. BOLT-LMM and SAIGE). If not, then please update this, as I am confident that your PGS results will improve.

6. You use SuSiE to finemap the variants, using LD derived from the imputation reference panel. I worry that this imputation panel might not capture the LD in the GWAS sample (with imputed SNPs). I believe you could improve the fine-mapping by

switching to the LD between the (imputed) SNPs in the GWAS sample, although this will likely not change much in practice.

7. The description of the novel variant identification procedure is unclear to me. You write (1) “We classified a gene or region as novel if the fine mapped independent signal was not located within 1 Mb of any reported genome-wide significant association (p-value < 5x10-8) for the corresponding phenotype55.” and (2) “Additionally, a variant was considered a novel hit if the highest linkage disequilibrium (LD) r^2 was less than 0.1 with any reported significant association.” Do you require both conditions, (1) and (2), to be true or is (2) enough to say the variant is novel?

8. You note that you use LDSC to estimate the heritability. However, I wonder how this estimate compares with the LDpred2-auto estimate, which assumes a point-normal prior, and could be much more accurate for outcomes with relatively simple genetic architectures, e.g. lipid measurements. Alternatively, you could use SBayesS to estimate the heritability.

9. It was unclear to me how you converted the r2 to liability scale. Was it using the same equation as the heritability, or something like Lee et al. (Gen Epi 2012)?

10. I am surprised that height has h2 of only 0.323 and BMI a h2 of 0.218. These seem small compared to the literature. Although this could very well be possible, it makes me wonder whether the GWAS was adjusted for sex. (Perhaps this relates to the following comment?)

11. In the validation of PRS, it seems that you don’t adjust for anything. Although you’ve already accounted for sex, age, chip, PCs, etc., in the GWAS, this can be an issue when comparing the r2 to h2, and when predicting into other cohorts. E.g., if you estimate h2 for height after adjusting for sex, then you should also adjust for sex when estimating the PGS r2, otherwise they won’t be on the same scale. Also, when using UKB-derived PGS to predict into TPMI, or the TPMI-derived PGS when predicting into other cohorts, it is possible that PCs could confound the prediction. In summary, I recommend that you always adjust for sex, age, PCs in the PGS analyses. (Interestingly, I found that the r2 is also reported with covariates in the supplementary tables.) For the AUC, you can perhaps compare it to an AUC of a base model (with standard covariates).

12. Please provide logistic regression or linear regression GWAS summary statistics online, (as well as SAIGE or BOLT-LMM). This would make it much easier for researchers to use these in future studies, as many common methods have issues with using mixed model summary statistics, e.g. LDSC, LDpred2, PRS-CS, lassosum2, MEGA-PRS, SBayesS, etc. This would therefore also increase the impact of the study.

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)
The revised manuscript has thoroughly and comprehensively addressed all my concerns through a more nuanced exposition of the results highlighting ascertainment bias and reducing the over interpretation of the results. As such, I find the revised manuscript greatly improved and I support its publication.

Referee #2

(Remarks to the Author)
I want to thank the authors for being responsive to feedback and improving their manuscript. I do not have any huge substantiative comments/concerns, but have listed some suggestions below to strengthen the manuscript, both in terms of framing and presentation.

Improving the narrative/flow of the piece:

The results would be strengthened by tying the TPMI results with the prevalence estimates either through NHIRD or through case numbers in UKBB. For example, in the section titled "cross-population comparison based on GWAS" (page 14, lines 534-542), it would have helped to tie the differential correlations with possible differences in the trait distributions. This would also influence the PRS scores, such as the higher performance of T2D (compared to SNP-heritability) compared to other traits.

The possible selection biases with TPMI also ties directly in with the last section on overall health measures. IT makes sense that cardiometabolic diseases are enriched in a clinical and hospital setting. This will also influence the statistical power and predictive ability of genetics. I would have liked to see this explicitly explored/mentioned, either in the results or discussion (or both to tie it all in throughout).

This is not necessary for publication, but I think it would be great to have at least one vignette throughout, such as HBV. So describe the prevalence, how it compares to population-level estimates. Then to discovery, genetic correlations, and PGS. And finally the contribution to overall health. That would make the piece seem more cohesive and give readers an example to point to. Right now there is so much (which is great!) but so much gets buried in the figures/supplement. A narrative is always appreciated! You do have this at the end for the discussion, but having all the results laid out, maybe in a box, would help set up this discussion better.

Minor presentation/word-choice comments:

Figure 1: Would have liked to see the distribution of correlations by trait type. For example, it looks like there are a cluster of respiratory traits that have a much lower prevalence in TPMI compared to the prevalence in NHIRD. This would make sense for something like asthma. This would also help give an idea of what types of outcomes are enriched in TPMI compared to NHIRD as those above the identity line. This further informs your genetics results, such as the lower replication for respiratory traits (which you do attribute to limited case numbers, but mentioning this further up explicitly would strengthen.)

Figure 2B would be stronger as a bar chart with the trait names above it.

My eyesight isn't great, so take this with a slight grain of salt, but please have mercy on my aging eyes and make all of the text in figures much larger/clearer. I would even make the lines/points larger. Even with zooming in, it was a challenge!

Line 676 calls HBV a population-unique disease. It is not unique to this population, but is enriched. Perhaps population-enriched instead of unique?

Line 715-716 states that genetics explains 10.3% of variation in hospital duration in Taiwan. It is more specifically in TPMI. Given the selection biases presented in the results, I would be cautious about extrapolating to the entire country.

Referee #3

(Remarks to the Author)
I would like to thank the authors for a very through response. I think the paper has improved substantially and very interesting! I found the comparison of hertiabilities between biobanks to be interesting. I'm stubbornly still surprised by this. Could this be due to an overestimated effective sample size in the LDSC, which would lead to downward bias? (probably not) Perhaps use a lower relatedness threshold? See perhaps also https://pmc.ncbi.nlm.nih.gov/articles/PMC10066905/ (which describes a related issue, but I don't think this is a plausible explanation here). Lastly, perhaps you can shed some light on this by expanding the LDpred2/SBayesS heritability estimates to more common outcomes. Those methods may have advantages over LDSC when it comes to dealing with varying SNP sample size issues. It would at least be interesting to see whether they are very different for, e.g. height and BMI. Another possible explanation is that height and BMI are subject to less assortative mating biases than in, e.g. the UK. (See e.g. Yengo et al. 2018, and others)

Version 2:

Reviewer comments:

Referee #3

(Remarks to the Author)
I would like to thank the authors for exploring the heritability estimation further. I think this is really interesting analysis. I would recommend the authors to include some of the analysis in the main manuscript or supplementary analysis, as it is in my opinion a very interesting and important finding. Also, the paper cited to argue the choice of LDSC, predates and does not cite LDpred2 or SBayesS. The benefit from using LDpred2 and SBayesS is that it estimates the heritability under a non-infinitesimal prior, meaning that the estimates should in theory be more accurate for outcomes that do not have an infinitesimal genetic architecture.

Referee #1:

The manuscript by Chen et al presents results of GWAS, heritability, genetic correlations, colocalization and PRS over a new biobank data of ~500k individuals within Taiwan (Taiwan Precision Medicine Initiative, TPMI) trying to identify population-specific genetic risk as a major source of risk in this population. The main addition to community is the resource of GWAS/PRS summary results over hundreds of ICD-based phenotypes that will likely provide another non-European centric genomic resource; I commend the authors for providing the summary results as open resource. As the community greatly needs more diverse biobank data, I am supportive towards the publication of this work. However, I find the majority of the claims being overstatements not fully supported by analyses presented and I encourage the authors to rephrase their results.

We thank the reviewer for recognizing the value of the GWAS/PRS results we generated from the TPMI dataset. As detailed below, we take the reviewer's concerns seriously and have addressed the perceived "overstatements" by rephrasing our results and conclusions as suggested. We believe that by addressing the important points raised by the reviewer, we have strengthened our manuscript and enhanced its scientific rigor.

1. Population labels. Terms such as "East Asian (EAS)", "Han Chinese", "Asian" need to be defined rigorously clearly labeling which ones are referring to genetically inferred ancestry versus self-reporting. For genetically inferred ancestries, it is important to clarify exactly what computational method used and how many individuals were excluded and did not meet the criteria. For self-reporting, it is important to describe the selection criteria for the inclusion in the biobank. The manuscript erroneously uses these different concepts interchangeably.

Response: We recognize the importance of distinguishing between genetically inferred ancestry and self-reported ancestry to ensure scientific precision and clarity when labeling populations. Below, we provide the necessary clarifications and revisions to address this concern.

Briefly, for the Taiwan Precision Medicine Initiative (TPMI) and Taiwan Biobank (TWB), all participants are referred to as "Han Chinese" based on genetically inferred ancestry. Genetic inference was conducted with principal component analysis (PCA), with data from the 1000 Genomes as reference panel, and ADMIXTURE was used to estimate the proportion of global ancestry. The detailed approaches are described below and have been incorporated into the supplemental document and main text accordingly:

Ancestry estimation

To determine the ancestral population of each sample, we used both principal component analysis (PCA) and ADMIXTURE, with data from the 1000 Genomes Project as the reference panel[1,2]. We included 88,695 linkage disequilibrium (LD)-pruned common variants (minor allele frequency, MAF > 0.05) across the five superpopulations of the 1000 Genomes Project: European (EUR), East Asian (EAS), South Asian (SAS), Admixed American (AMR), and African (AFR). The PCA was conducted with unrelated subjects from the 1000 Genomes Project, and then all TPMI genotyped samples were projected based on the weightings from 1000 Genomes PCA (Fig. 4 in the revised supplementary document). The first two principal components separated EAS from the other superpopulations, and 168 subjects were excluded due to their genetic distance from the EAS cluster. We then applied ADMIXTURE to quantify the proportion of genetic ancestry among EAS, resulting in the inclusion of 161,194 individuals from TPMv1 and 328,307 individuals from TPMv2 with high genetic similarity to Han Chinese in further analyses (Table 1 and 2).

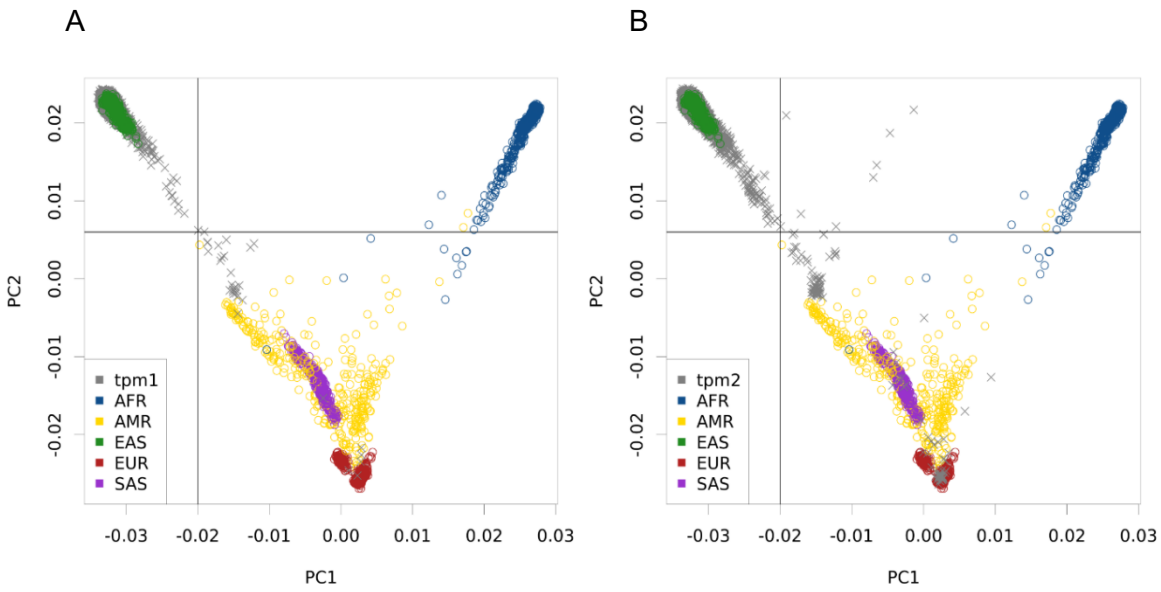


Figure R1-1. (Figure 4. in the revised supplementary document) PCA results for individuals from 1000 genomes and TPMI genotyped withTPMv1 (A) and TPMv2 (B)

Table 1. ADMIXTURE result for individuals genotyped with TPMv1

Est. proportion	v1	v2	v3	v4	v5	-
CDX	0.041	0.014	0.839	0.081	0.025	
CHB	0.264	0.018	0.123	0.571	0.024	
CHS	0.132	0.025	0.286	0.509	0.049	
JPT	0.916	0.007	0.016	0.051	0.009	
KHV	0.083	0.018	0.727	0.145	0.027	
	JPT	TW2*	CDX + KHV	**Han**	TW1*	NOS
TPMv1 (n=168,330)	27	1,228	391	**161,194**	2,741	2,749

Table 2. ADMIXTURE result for individuals genotyped with TPMv2

Est. proportion	v1	v2	v3	v4	v5	
CDX	0.011	0.858	0.045	0.020	0.066	
CHB	0.016	0.135	0.255	0.024	0.569	
CHS	0.022	0.308	0.130	0.047	0.492	
JPT	0.007	0.016	0.918	0.010	0.049	
KHV	0.016	0.741	0.084	0.023	0.137	
	TW2*	CDX + KHV	JPT	TW1*	**Han**	NOS
TPMv2 (n=340,401)	936	1,396	99	2,731	**328,307**	6,932

* The ancestral populations in Taiwan other than Han Chinese

The above information can be found in the supplementary document (p. 10-11).

For the UK Biobank (UKB) analysis, ancestry was based on self-reported ethnicity. We used only self-reported Chinese (more diverse than Han Chinese) as East Asian for external validation. For the All of Us analysis, we used their genetically confirmed ancestry based on PCA and ADMIXTURE with 1000 Genomes as reference. To ensure transparency, ancestry is labeled as "self-reported" unless additional genetic validation is explicitly described in the respective dataset methodologies.

Revised sentences (in results, LINE 404-407)

We performed comprehensive genomic analyses, including GWAS, heritability estimation, and PRS model building and evaluation, across a wide range of diseases and quantitative traits using 463,447 genetically inferred Han Chinese from TPMI.

Revised sentences (in results, LINE 588-593)

To evaluate the robustness and generalizability of our PRS models, we performed an external validation of the models (hypertension, type 2 diabetes, viral hepatitis B, gout, calculus of kidney from PRSmix+ and others from LDpred2) in Taiwan Biobank (TWB, genetically inferred unrelated Han Chinese, n=88,628), UKB (self-reported EAS, n=9,893), and All of Us (genetically inferred EAS, n=6,895).

A detailed description of the definition of EAS was incorporated into the method section in the revised manuscript.

Revised sentences (in methods, LINE 1005-1020)

UKB has enrolled ~500,000 participants since 2006 and linked their genetic data with enriched phenotypic data. For UKB validation, we used their inpatient record for case definition. Their ancestral population was determined by self-reported ethnic background, such as self-reported Chinese as EAS (n=1,572), White, British, Irish, and any other white background as EUR (n=472,869), Black or Black British, Caribbean, African, and any other Black background as AFR (n=8,074), and Asian or Asian British, Indian, Pakistani, Bangladeshi, and any other Asian background as SAS (n=9,893). All of Us intends to enroll over 1 million participants in the United States and has released whole genome genotyping data for ~312,000 participants as of the first quarter of 2024. We applied ADMIXTURE with 1000 Genomes data as reference panel for assigning the genetically inferred ancestral populations, including EAS (n= 6,895), EUR (n= 152,754), AFR (n= 60,964), AMR (n= 32,394), and SAS (n=2,334). The genetically confirmed EAS as well as other superpopulations and their linked EMR were used for validating our PRS models.

2. TPMI Biobank representativity of Taiwan. The authors compare disease rates in the biobank versus the National Health Insurance Research Database in Taiwan(NHIRD) finding a correlation in the prevalences of r=0.64. I disagree with their conclusion that a correlation of 0.64 is proof that that TPMI represents well the population. Such a small correlation could be indicative of ascertainment issues and/or problematic phenotypic based on phecodes. Moreover, the case proportion of TPMI seems deflated across the board as compared to NHIRD prevalence (Figure 1a) thus suggesting that TPMI samples a healthier snapshot of the population. The authors could look at distributions of age/sex/locations/etc to describe the main drivers of ascertainment in TPMI. Even if TPMI is not a perfect representation of general population (which it seems to be the case), it is still a great resource and highly useful for discoveries. I suggest the authors describe this in the text of section discussing Figure 1.

Response: We re-calculated the correlation with several previously missing NHIRD prevalences, and the updated coefficient is r = 0.656, with a significant p-value 2.69×10^{-84}. We also include the

NHIRD prevalence in the Table S1 now. We acknowledge that the correlation of 0.656 between TPMI and NHIRD disease prevalence indicates a moderate alignment rather than full population representativeness. To ensure a more precise interpretation of this result, we have explicitly discussed ascertainment biases and demographic differences in the revised manuscript.

To further explore potential ascertainment biases, we analyzed the demographic distributions of TPMI participants and compared them to NHIRD. We found that TPMI participants are overrepresented in the middle-aged group (year of birth 1941-1970), comprising 54.3% of the cohort, compared to 38.3% in the NHIRD. Additionally, TPMI includes a slightly higher proportion of female participants (55.1%) compared to NHIRD (50.6%). (Figure R1-2) Furthermore, TPMI has a higher proportion of participants from northern Taiwan, with 59.5% recruited from hospitals located in northern Taiwan, compared to 47.3% in the NHIRD from individuals residing in northern Taiwan. in contrast, less TPMI participants are from southern Taiwan (10.3% vs. 24.9%), These demographic differences, combined with TPMI's volunteer-based recruitment process, likely contribute to the lower proportion of cases observed in TPMI compared to NHIRD, as illustrated in Figure R1-3.

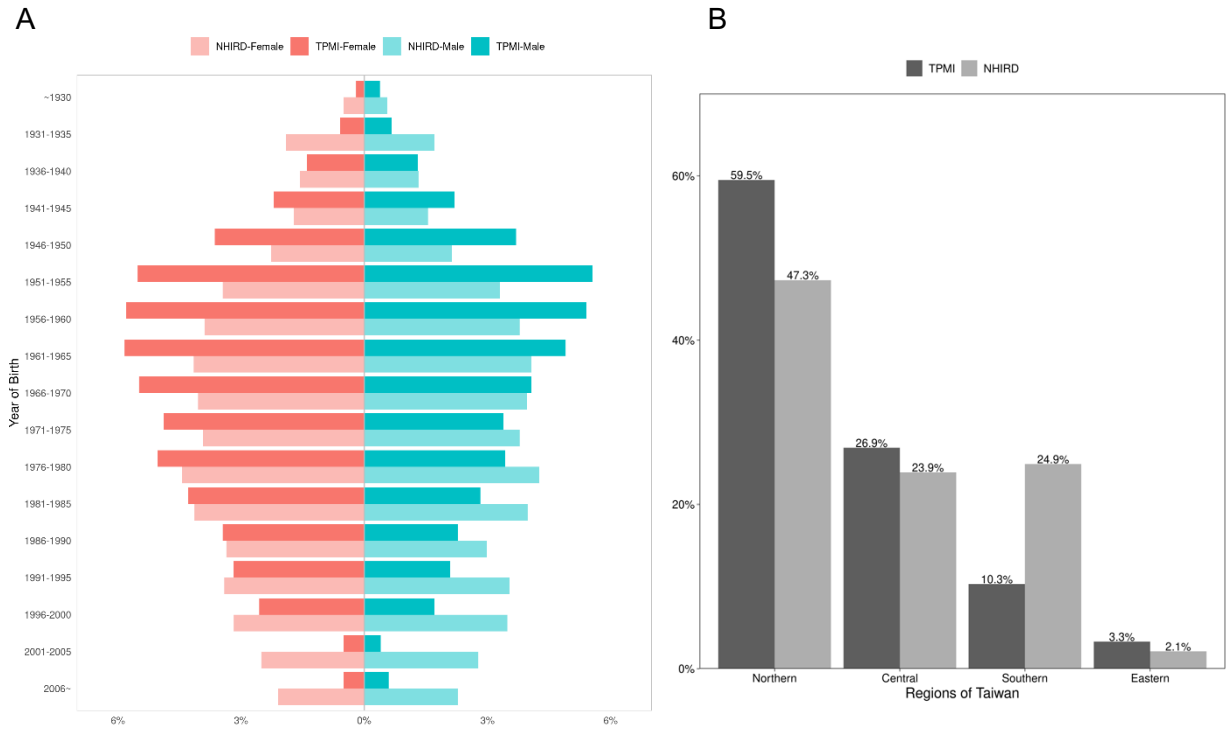


Figure R1-2. The basic demographic information comparison between TPMI and NHIRD for year of birth by sex (A) and regions (B)

Additionally, we recognize that TPMI primarily recruits participants from medical centers, whereas NHIRD includes data from a broader range of healthcare providers, including local clinics and primary care settings. This difference in healthcare setting coverage may lead to an underrepresentation of milder and more common diseases in TPMI, as these conditions are often diagnosed and treated in local clinics rather than hospitals. The exclusion of local clinic data from TPMI's electronic medical records (EMRs) further contributes to discrepancies in disease

prevalence between TPMI and NHIRD. We have highlighted these strengths and limitations more explicitly in the revised manuscript.

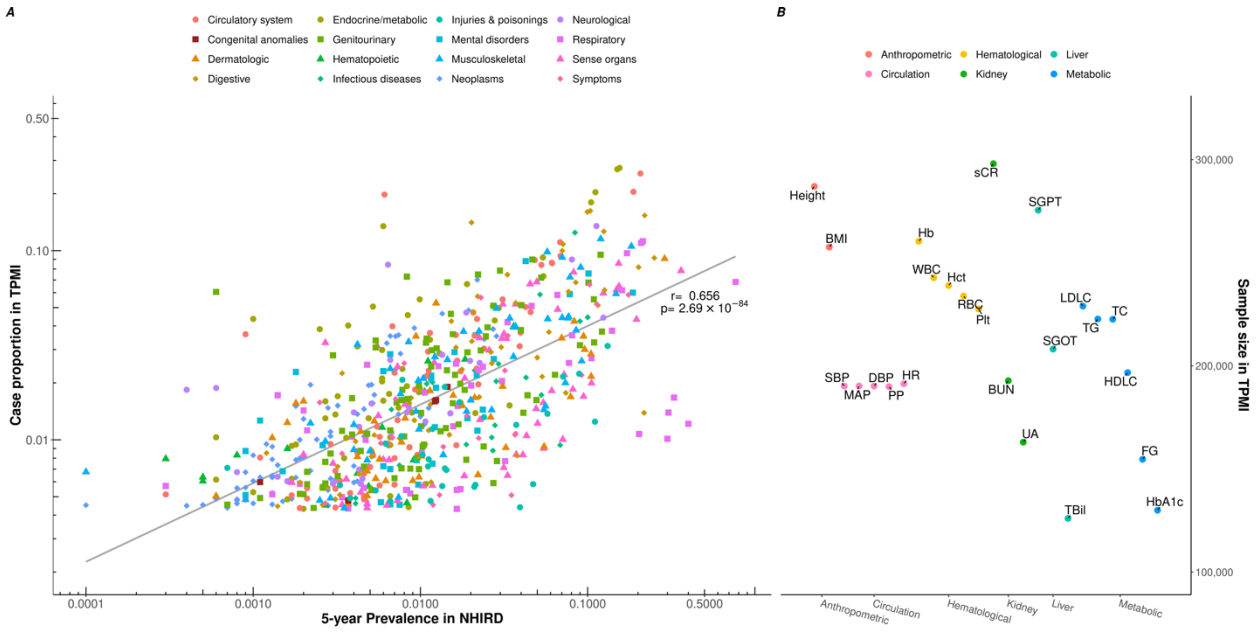


Figure R1-3. (Figure 1 in the revised manuscript) Scatter plots of the case proportion and case number in TPMI dataset. (A) The case proportion in TPMI is compared to the 5-year prevalence in the National Health Insurance Research Database (NHIRD) for dichotomized phenotypes (phecodes). Each dot represents a specific phecode, with the x-axis showing the prevalence in NHIRD and the y-axis showing the case proportion in TPMI. (B) Scatter plot shows the sample sizes for quantitative traits in the TPMI cohort. Each point represents a trait, with the x-axis indicating the different category of quantitative traits and the y-axis representing the corresponding sample sizes.

Revised sentences (in results, LINE 414-420)

The log-transformed case proportion identified from EMR showed an imperfect but significant correlation with the log-transformed 5-year disease prevalence from National Health Insurance Research Database in Taiwan[3] (r = 0.656, p-value = 2.69×10^{-84}) (Fig. 1A), suggesting that the TPMI's hospital-based design may not fully capture mild and common illness, which are primarily observed in local clinics and primary care.

And we also discussed this assessment bias in the discussion.

Revised sentences (in discussion, LINE 722-746)

As other large-scale epidemiological studies, ascertainment bias is also observed in our study. TPMI's case proportion also shows a significant but imperfect correlation with the prevalence from National Health Insurance dataset, and this may imply the ascertainment bias of TPMI's hospital-based design. Compared to the general population in Taiwan (NHIRD), TPMI participants are overrepresented in the middle-aged group (year of birth 1940-1970, 54.3% vs. 38.3%), include slightly more females (55.1% vs. 50.6%), and have a higher proportion of participants from

northern Taiwan (59.5% vs. 47.3%). These demographic differences, along with the volunteer-based recruitment process, likely contribute to the lower case proportions observed in TPMI (Figure 1a). Importantly, a significant portion of disease records in the NHIRD originate from local clinics and primary care settings, which are not covered in TPMI. Notably, the EMR of the participants are incomplete, as some participants receive care from multiple health providers, but the TPMI only has access to EMRs from their enrollment hospitals. We acknowledge that ascertainment biases may influence disease prevalence and heritability estimates. Thus, we accounted for case-control ascertainment in heritability estimates and PRS evaluations by applying liability-scale transformations using population prevalence data from NHIRD and utilized independent validation datasets for PRS validation (TWB, UKB, All of Us) to mitigate the impact of ascertainment biases, while acknowledging that residual biases may persist. Methods like inverse probability weighting could mitigate such biases[4], but these require detailed external reference data that are currently unavailable. This limitation also emphasizes the need for future adjustments to enhance generalizability.

3. A central theme of the work is that population-specific genetic risk is a major driver for complex traits. I find this claim not supported in the data. For example, cross-population correlations are very high for most traits (Table S6) (lines 508-514). Furthermore, the vast majority of the loci have been reported before (only 77 out of 1,139 are novel). Can the authors report which loci have evidence of heterogeneity across populations? How many of those heterogeneities can be ascribed to true differences in genetic architectures (as claimed on lines 513-514) and how many are due to LD-tagging or differences in frequencies?

Response: We apologize for missing the bug in our counting script, which resulted in an undercount of the lead variant in the failed fine-mapping region of the GWAS with successful fine-mapping regions. After correcting this error, we now report 2656 fine-mapped independent variants, including 1,309 variants from the phecode GWAS and 1,347 variants from quantitative traits. To address the reviewer's concern, we conducted additional analyses and incorporated the results into the revised manuscript. First, we compared the allele frequencies of the novel variants identified in TPMI with those observed in European populations. Among the 95 novel loci, 33 have an allele frequency below 0.01 in European populations, making them difficult to detect in European-based GWAS. This result supports the idea that certain disease-associated variants may be unique to the Han Chinese or have a much lower prevalence in non-East Asian populations. Second, we assessed genetic heterogeneity of genetic effect on the target traits between TPMI and UKB using t-test. The results suggest 25 loci have significantly distinct effect size between TPMI and UKB. These results provide further evidence of population-specific effects in disease susceptibility. The full results of these analyses have been incorporated into Table S5 of the revised manuscript. Third, we identified 217 new signals located within previously reported genomic regions but with low linkage disequilibrium (LD) ($r^2 < 0.1$) with known variants, called new hits in our revised document. This finding suggests that while these signals are located in regions previously implicated in disease risk, they represent distinct population-specific variants that are likely to influence disease susceptibility through different genetic mechanisms. The observed differences further highlight the genetic heterogeneity across populations and

underscore the importance of studying diverse populations to improve the generalizability of genetic findings.

To ensure that these findings are clearly communicated, we have revised the Results section to explicitly discuss these novel findings and their implications.

Revised sentences (in results, LINE 450-471)

Of the 95 novel genetic associations, 30 variants are rarely found (MAF<0.05) in other populations (African, Admixed American, South Asian, and European (non-Finnish) in gnomAD), and 33 variants less than 0.01 in the Europeans, the most extensively studied population, which explains why they were not reported in previous GWAS. For example, the SNP rs17089782, a missense variant in the *PIBF1* (p.R405Q) gene on chromosome 13 is significantly associated with thyroid cancer ($p = 2.8x10^{-9}$) in the TPMI cohort. This SNP, rs17089782, has an allele frequency of 5.65% in TPMI but 0.01% in European populations. This population-specific difference likely explains why this association was detectable in TPMI but not in predominantly European datasets. However, *PIBF1* is essential for immune regulation, especially during pregnancy, and is relevant to autoimmune diseases and cancer[5]. Another novel variant identified our quantitative analysis of BMI (rs761018157 p-value = $4.8x10^{-9}$, MAF in TPMI = 4.34%, MAF in EUR<0.01%) maps to *PHOX2B*. This gene, highly expressed in the nervous system, had previously been linked to obesity hypoventilation syndrome in a small study (n = 30)[6] and associated with bone mineral density[7]. In addition, when we compare the effect size in TPMI and UKB for the rest of the novel findings, 25 exhibit a significant different effect size (p-value<0.05). For instance, a TPMI-identified platelet count-relevant variant, rs12955741, located in the intergenic region between *TGIF1* and *DLGAP1* exhibits a different effect size compared to that in UKB ($\beta_{TPMI} = 0.044$, $\beta_{UKB} = -0.005$, p-value = $1.7x10^{-9}$).

4. GWAS sample size. Methods section (line 752) suggests that GWAS was performed after excluding 100k unrelated individuals; I found that problematic in unnecessarily restricting power. Moreover, line 764 notes that only 248k (out of 500k) subjects are "unrelated"; it is highly unusual to have so many related samples in the biobank. Can the authors clarify sample sizes for each analysis? Can the authors clarify excluded sample sizes based on various criteria? It seems like 20k subjects were excluded based on quality control alone (line 729).

Line 728-730: As a result, 401,710 genetic variants and 463,447 Han Chinese participants passed all quality control measures and were used in the subsequent studies.

Line 752-753: To train and validate the proposed PRS models, we left 100,000 unrelated subjects out of GWAS.

Line 762-765: In addition, we also performed a generalized linear model from PLINK2 for GWAS among the unrelated subset (n = 248,754), and these statistics were used for heritability and genetic correlation estimation.

Response: To develop a reliable PRS model, we followed the suggested pipeline from previous published papers[8-10]. First, we conducted a GWAS to prioritize genetic variants, then used independent individual-level data to fine-tune the weights and parameters of the PRS model and evaluated its performance with another independent set. The overlap or reuse of samples between any steps could lead to model overfitting or inflation. Therefore, the unrelated set we

mentation (n=248,754) was extracted from the GWAS set (n=363,477). To clarify this confusion, we have revised the sentences in methods.

Revised sentences (in methods, LINE 866-881)

The entire dataset was divided into three subgroups: the GWAS set (N = 363,447), the training set (N = 80,000), and the testing set (N = 20,000). …… Within the GWAS set, we selected an unrelated subset (N = 248,754) to perform GWAS using a generalized linear model in PLINK2and we conducted 1:10 age, sex-matching for the traits with imbalanced case/control ratio (<1/20). These GWAS statistics were then used for heritability and genetic correlation estimation.

Alos, we have a table listing the number of subjects excluded during the QC process in the supplemental document (page 13) for clarification. The majority of excluded subjects were from non-Han Chinese ancestry or without EMR.

Table 3 Summary table for genotyping QC

	TPMv1		TPMv2		Combined	
	Subjects	Variants	Subjects	Variants	Subjects	Variants
Raw number	**168,375**	**686,439**	**340,537**	**725,699**		
Step 1 per-batch filtering						
Missing rate by batch		-145,132		-50,955		
After QC by Batch (intersect)	168,375	541,307	340,537	674,744		
SNP-level missingness difference		-4,686		-10,137		
Difference of allele frequency between batches		-5		-231		
After stage 1 filtering	**168,375**	**536,616**	**340,537**	**664,376**		
Step2 merged-batch filtering						
Remove disqualified variants		-704		-2		
Remove duplicate variants		-73		0		
Remove samples > 5% missing	-12		-1	0		
Remove variants > 2% missing		-112		-89		
After stage 2 filtering	**168,363**	**535,727**	**340,536**	**664,285**		
Step 3 ancestry determination						
East Asian	168,330	535,727	340,401	664,285		
Han Chinese	**161,194**	535,727	**328,307**	664,285		
Step 4 post-ancestry filtering						
Remove variants > 2% missing		-3		0		
Remove rare variants (MAF<1%)		-63,730		-26,231		
Check sex (F coefficient)	-336		-530			
Opposite sex with EHR	-240		-431			
Heterozygosity filter	-6		-6			
Remove rare variants (MAF<1%)		-34		-30		
After Sex/Het checking	160,612	471,960	327,340	638,024		
Non-autosome		-11,093		-21,672		
HWE <1e-6		-16,662		-21,759		
After Autosome/HWE	160,612	444,205	327,340	594,593		
Batch GWAS		-947		-9,440		
Multiallelic		-741		-477		
Han after QC by Array	**160,612**	**442,517**	**327,340**	**584,676**	**487,952**	**401,777**
Step 5 Array merging						
Effect of array version						-67
Duplicate					-4,299	
Han after QC					**483,653**	**401,710**
Han with EMR (EMR n=500,081)					**463,447**	**401,710**

5. Comparison with other Taiwanese biobank (TWB). The authors do a limited comparison with TWB when discussing external validation of PRS models. Since TWB has similar sample size as TPMI, it would greatly enhance the manuscript if the authors also compare the GWAS/h2/correlations across the two biobanks. Furthermore, if the manuscript presents PRS that would be the best for Taiwanese individuals, it seems like combining results across the biobanks would surely provide superior accuracy than just one biobank.

Response:

While this would be ideal, it is currently not feasible due to key differences in data structure and quality. Specifically, TWB has over 200,000 collected samples, but only ~150,000 genotyped datasets are currently available, and these rely primarily on self-reported phenotypes without ICD coding or detailed medical records, which decreases significantly the number of traits in our current medical phenome study. In contrast, TPMI includes over 500,000 participants with both genotyping data and comprehensive medical records-based phenotypes. These differences in phenotype definitions and clinical data availability pose significant challenges for direct GWAS, heritability, or PRS comparisons, as they could introduce biases or inconsistencies.

We have discussed potential improvement by integrating other large-scale EAS biobanks. However, work is underway to combine the TWB data and build a compatible, robust combined TPMI/TWB dataset.

Revised sentences (in discussion, LINE 770-774)

Furthermore, the meta-analysis integrating TPMI with other large-scale EAS biobanks, such as Taiwan Biobank, Biobank Japan, Korean Genome and Epidemiology Study, and China Kadoorie Biobank, may further enhance our understanding of the genetic etiology in EAS and improve prediction models.

6. TPMI-derived PRS outperformed UKB derived PRS when applied to EAS ancestries; this is not surprising and great to see. But wouldn't that suggest that a joint model combining EUR_PRS with EAS_PRS would be the best?

Response: To address this, we have included Figure R1-4, comparing AUC values for PRS models derived from TPMI alone and TPMI+UKB with PRS-CSx.

As shown in Figure R1-4, PRS models trained solely in TPMI perform similarly or slightly better for prostate cancer, viral hepatitis B, and type 2 diabetes, while TPMI+UKB-derived models show comparable or improved performance for alcoholism, glaucoma, and migraine. These results highlight population-specific differences in PRS performance and emphasize the importance of ancestry-matched training data. We have incorporated the cross-population PRS into our results and, we suggest this would be a further direction for the PRS development and application.

Revised sentences (in results, LINE 612-619)

Additionally, we assessed the performance of TPMI-derived PRS and TPMI-included cross-population PRS across various ancestry groups, including European, African, Admixed American, and South Asian populations from the UK Biobank and All of Us cohorts (Extended Data Fig. 6). The performance varied by diseases, but consistent results were observed for female breast

cancer and glaucoma across populations, and TPMI-included cross-population PRS slightly but not significantly improved in the populations other than EAS and EUR.

The Discussion section has been expanded to address PRS transferability across populations.

Revised sentences (in discussion, LINE 707-714)

When comparing with European-derived PRS models, we also observed better performance across several traits in EAS, particularly for cardiometabolic and autoimmune diseases. Similarly, the TPMI-included cross-population model slightly improves performance in other ancestral populations. These results highlight the need for population-specific prediction models and emphasize the importance of genetic data from diverse populations to advance cross-population models.

7. The authors find that PRS for various traits correlate with number of visits and hospitalization

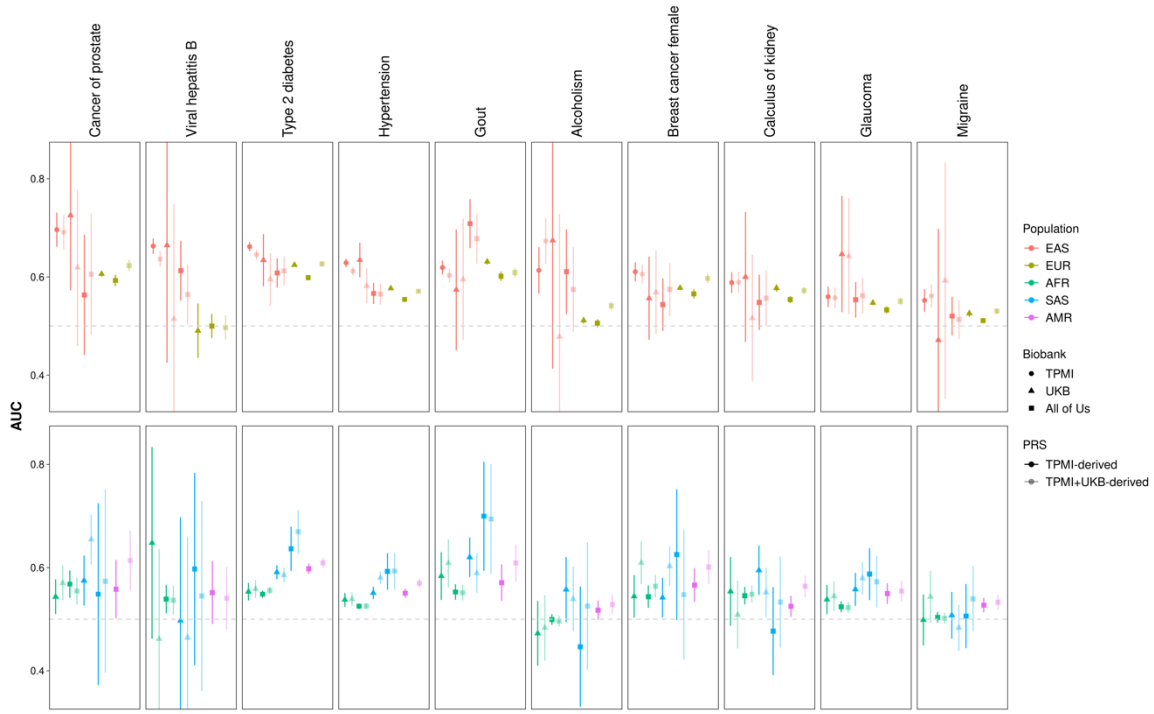


Figure R1-4. (Extended Data Fig. 6 in revised manuscript) External validation of TPMI-derived PRS model and TPMI-UKB cross-population PRS models across populations. PRS validation (AUC ± 95% CI) in East Asian (red), European (olive green), African (green), South Asian (blue), and All of Us (purple) populations from TPMI (circle), UKB (triangle), and All of Us (square) cohorts.

duration explaining ~9% of the variation. Part of this could be due to non-genetic factors due to residual stratification in the PRS. How much are other factors like age, sex, hospital location, explaining in this trait? Were they adjusted for in the analyses of lines 572-578? This may impact the statement of lines 643-644 as non-genetic factors may drive predictability. Also important to note in text that number of visits and/or days in the hospital are a poor pricy for "disease burden". Also note that TPMI may not be a perfect proxy for all Taiwan (see comment above).

Response: We acknowledge that residual stratification in PRS and non-genetic factors such as age, sex, and hospital location may contribute to the observed association between PRS and variation in overall health indices metrics. These factors have been adjusted for in the analyses

on lines 572–578. To clarify this, we have clarified them the manuscript to include their contributions in the supplementary tables (Table S15). As suggested, we have revised the text to note that the number of visits and hospitalization duration are imperfect proxies for "disease burden".

Table R1-4. (Table S15) The partial explained variance (R^2) of covariates in the health indices models.

		Top 5% vs Bottom 5%		Top 10% vs Bottom 10%		Top 20% vs Bottom 20%	
Outcome	Covariate	r^2	p-value	r^2	p-value	r^2	p-value
Clinical Visit	Age	0.138	2.63E-49	0.129	1.59E-90	0.110	1.22E-151
	Sex	0.000	4.56E-01	0.000	9.02E-01	0.000	6.85E-01
	Hospital	0.438	1.45E-165	0.397	7.91E-299	0.336	<1E-300
Hospitalization	Age	0.000	4.77E-01	0.000	4.84E-01	0.000	4.66E-01
	Sex	0.000	9.00E-01	0.000	5.63E-01	0.000	7.57E-01
	Hospital	0.002	9.65E-01	0.001	9.69E-01	0.002	3.05E-01

Revised sentences (in results, LINE 621-632)

Although overall health is hard to define with a few metrics, herein, we used the count of clinical visits and duration of hospitalization to roughly describe individuals' overall health. We found that 131 of the top performing PRS models (LDpred2 models with AUC > 0.55 and all PRSmix+ models for phecodes and all models for quantitative traits) are significantly associated with overall health indices, explaining 8.47% of the variation in clinical visit frequency (p-value = 2.69×10^{-14}) and 10.29% of the variation in hospitalization duration (p-value = 5.62×10^{-27}, Table 1 and S15) in the comparison between top and bottom 5% groups after adjusting sex, age and recruiting hospital. Among the identified clusters, the cardiometabolic disease cluster contributed the most to the indices, accounting for 1.32% of clinical visits (p-value = 0.02) and 3.55% of hospitalizations (p-value = 7.10×10^{-9}).

Additionally, we have highlighted that TPMI may not fully represent the entire Taiwanese population due to potential ascertainment bias, as discussed above

8. GWAS replication rates range from 19.81% to 74.4%. To what do the authors ascribe this? A replication rate of 10-30% seems extremely low. Can the authors stratify replication rates according to statistical power expected (as function of frequency and effect size in existing work)? I find the statements on lines 438-439 not assuring at all.

Line 436-439: Lower replication rates for respiratory and psychiatric disorders (27.78% and 19.81%) may reflect limited case numbers, other untyped genetic variants, such as rare variant, copy number variation and structure variants, or recruitment bias.

Response: We now report the replication rate with the actual over expected ratio (AER), where the expected number of replicable associations was determined after considering power, and updated the extended Figure 1. When we only consider the associations where we have sufficient power (>0.8), the replication rate for all phecodes would increase from 74.4% to 75.3%. The two lowest groups would show improvements, with the psychiatric group increasing from 19.81% to 25.00%, and the respiratory group holding steady at 27.78%. We have now included both the raw

replication rate, the stratified replication rate (power > 0.8), and the actual/expected ratio in revised Table S3.

For the psychiatric group, the majority of reported associations are from schizophrenia (16/17), for which we have just over a thousand cases in our GWAS. After removing schizophrenia GWAS, we can replicate the only reported association in psychiatric diseases in EAS (alcoholism). The respiratory group is another underperforming trait, with asthma contributing the most reported associations in this group (17/18). However, we were only able to replicate four associations among them. Given the age distribution of the TPMI participants (very few under 18, Fig. R1-1A) [11], we may include mostly asthma patients exposed to environmental factors like air pollution and smoking, rather than those associated with genetic risk factors. The possible factors responsible for the low replication rate has been elaborated more in revised manuscript.

Revised sentences (in results, LINE 424-433)

Our GWAS identified at least one significant locus (p-value < $5x10^{-8}$) for 265 phecodes of the 695 tested and all 24 quantitative traits. Highlighting the robustness of the TPMI data, we observed a high replication rate of reported disease loci from EAS GWAS on GWAS catalog (actual/expected ratio (AER) = 78.17%, considering the statistical power with the published tool, PGRM[21]), particularly for endocrine and metabolic/hematopoietic diseases (AER = 88.68% and 84.62%, respectively; Extended Data Fig. 1 and Table S3). Lower replication rates for respiratory (AER = 23.53%) may reflect limited case numbers, untyped genetic variants, such as rare variants, copy number variation, and structural variants, or recruitment bias such as age distribution.

9. Lines 440-446 report that only 77 (out of 1,139) are novel associations with text noting how these highlight the uniqueness of these loci in TPMI (lines 457-459). Which of the 77 have support for heterogeneity (i.e. different causal effect) or are simply due different statistical power for discovery?

Lines 440-446: (Genome-Wide Association Studies, Fine-Mapping, and Novel Finding) Our GWAS revealed 1,139 independent association signals for phecodes and 1,305 signals for quantitative traits via fine-mapping. Notably, 77 novel associations were identified across 44 phecodes and 7 quantitative traits that had not been previously reported in nearby regions (±1Mb), and 307 novel hits from reported regions (Table S4 and S5). Some of our novel findings are biologically relevant to the corresponding phenotypes. For example, the SNP rs17089782, a missense variant in the *PIBF1* (p.R405Q) gene on chromosome 13 is significantly associated with thyroid cancer (p = $2.8x10^{-9}$).

Response to reviewer:

To determine whether the 95 novel loci identified in TPMI reflect true genetic heterogeneity or differences in statistical power, we conducted additional analyses and incorporated the results into the revised manuscript.

First, we examined genetic heterogeneity comparing the effect size from TPMI and UKB. These analyses identified 25 loci that exhibit significant heterogeneity across populations, suggesting that their effects differ between ancestries. One example is rs17089782 in *PIBF1*, which is

significantly associated with thyroid cancer ($p = 2.8\times10^{-9}$) in TPMI. The observed heterogeneity for this variant is consistent with immune-related genetic differences across populations. These results indicate that some genetic variants have ancestry-specific effects on disease susceptibility, reinforcing the need for population-specific GWAS. Second, we investigated whether some of the 95 novel loci were detected due to differences in allele frequencies. For 33 loci, these associations were not reported previously or genome-wide significant due to lower allele frequencies in European datasets. The large sample size and phenotype resolution in TPMI allowed us to detect these associations, which were previously undetectable in smaller or non-hospital-based datasets.

These results have been incorporated into Table S5 of the revised manuscript and are explicitly discussed in the Results section. Additionally, the Discussion section has been updated to highlight the implications of population-specific discoveries for PRS development and cross-ancestry risk prediction.

Revised sentences (in results, LINE 450-471)

Of the 95 novel genetic associations, 30 variants are rarely found (MAF<0.05) in other populations (African, Admixed American, South Asian, and European (non-Finnish) in gnomAD), and 33 variants less than 0.01 in the Europeans, the most extensively studied population, which explains why they were not reported in previous GWAS. For example, the SNP rs17089782, a missense variant in the *PIBF1* (p.R405Q) gene on chromosome 13 is significantly associated with thyroid cancer ($p = 2.8x10^{-9}$) in the TPMI cohort. This SNP, rs17089782, has an allele frequency of 5.65% in TPMI but 0.01% in European populations. This population-specific difference likely explains why this association was detectable in TPMI but not in predominantly European datasets. However, *PIBF1* is essential for immune regulation, especially during pregnancy, and is relevant to autoimmune diseases and cancer[5]. Another novel variant identified our quantitative analysis of BMI (rs761018157 p-value = $4.8x10^{-9}$, MAF in TPMI = 4.34%, MAF in EUR<0.01%) maps to *PHOX2B*. This gene, highly expressed in the nervous system, had previously been linked to obesity hypoventilation syndrome in a small study (n = 30)[6] and associated with bone mineral density[7]. In addition, when we compare the effect size in TPMI and UKB for the rest of the novel findings, 25 exhibit a significant different effect size (p-value<0.05). For instance, a TPMI-identified platelet count-relevant variant, rs12955741, located in the intergenic region between *TGIF1* and *DLGAP1* exhibits a different effect size compared to that in UKB ($\beta_{TPMI} = 0.044$, $\beta_{UKB} = -0.005$, p-value = $1.7x10^{-9}$).

10. Heritability estimates seem too low and different from other reported results (lines 468-476). First, the authors should clarify they report SNP-heritability results (not narrow-sense results). Second, how do these results compare to other biobanks? For example h2-snp of 0.3/0.2 for height/bmi seems too low. Since TPMI is not the only large-biobank from Taiwan, the authors could compare heritability results across biobanks.

Lines 468-476: (Genome- and Gene-level Heritability, and Colocalization) Linkage disequilibrium score regression analysis showed strong liability-scaled SNP-heritability for conditions such as alcoholism ($h^2 = 0.242$), open-angle glaucoma ($h^2 = 0.171$), and retention of urine ($h^2 = 0.173$). In terms of quantitative traits, body height ($h^2 = 0.323$), BMI ($h^2 = 0.218$), and high-density lipoprotein cholesterol ($h^2 = 0.191$) exhibited the highest heritability estimates (Table

S6), highlighting the significant role of genetics in these traits. These results have far-reaching implications for precision medicine, as higher heritability signals suggest the potential for more accurate genetic risk prediction models that could improve personalized disease risk assessments.

Response: To clarify, the reported heritability values represent SNP-heritability (h^2-SNP) estimated using Linkage Disequilibrium Score Regression (LDSC) and do not reflect narrow-sense heritability (h^2), which includes contributions from all genetic variation sources.

We compared our SNP-heritability estimates with those from Taiwan Biobank (TWB), Biobank Japan (BBJ), and UK Biobank (UKB), all of which applied the same statistical framework[12]. As shown in Figure R1-5, TPMI-derived heritability estimates are largely consistent with previous reports across multiple traits.

Although the SNP-heritability of height (h^2 = 0.323) in TPMI is slightly lower than that in Taiwan Biobank, this discrepancy is likely due to differences in phenotype measurement protocols. Specifically, TPMI height values are extracted from electronic medical records (EMR), which include both self-reported and measured values, while Taiwan Biobank follows a standardized direct measurement protocol. This variability may introduce additional noise, leading to slightly lower heritability estimates in TPMI.

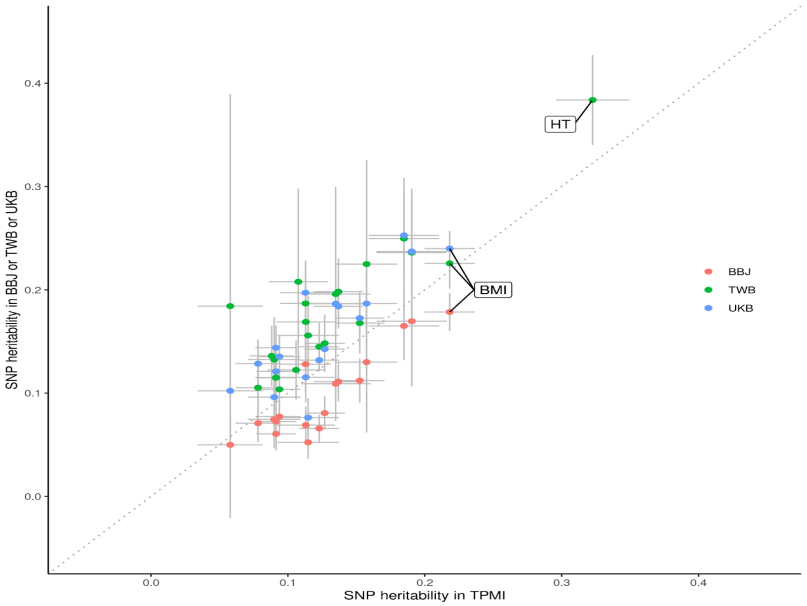


Figure R1-5. The estimates of heritability for 24 quantitative traits in TPMI (x-axis) and other biobanks (y-axis), including Taiwan Biobank (TWB, green), Biobank Japan (red), and UK Biobank (blue).

11. lines 487-490: unclear and too broad of a statement.

Lines 487-490: (Genome- and Gene-level Heritability, and Colocalization) These findings highlight the pleiotropic effects of these genes and present potential avenues for cross-disease therapeutic strategies, where targeting one gene could influence multiple related disorders. Understanding these shared genetic effects is crucial for devising broader precision medicine approaches, particularly for managing comorbidities.

Response: We have revised the statement to make the statement more specific.

Revised sentences (in results, LINE 510-520)

Our findings demonstrate the effect of these genes (such as *APOE, ABCG2*, and *KCNQ1*) on multiple traits and disorders (Table S6). For example, *APOE* influences both lipid metabolism and Alzheimer's disease risk, making it a promising target for therapeutic strategies aimed at managing these interconnected conditions. By elucidating shared genetic effects, these results offer opportunities to develop precision medicine approaches that address comorbidities, such as treating hyperlipidemia and reducing dementia risk through a single intervention targeting *APOE.* This gene-level understanding emphasizes the potential to optimize therapeutic strategies by leveraging genetic pleiotropy in disease management.

12. lines 501-506: how are these findings significant for clinical applications?

Lines 501-506: (Genetic Correlation and Clusters) These findings are significant for clinical applications, suggesting that shared genetic risks across diseases could enable earlier detection of comorbidities and prevention strategies based on an individual's genetic profile. The shared genetic architecture also implied that we may leverage the genetic risk of correlated traits while developing the PRS model.

Response: To clarify and highlight the clinical significance of our findings, we have rewritten those sentences and moved them into discussion as shown below

Revised sentences (in discussion, LINE 682-698)

In addition, the comprehensive phenotypic data allows us to leverage the genetic correlation of multiple traits to improve the performance of PRS models. Three large clusters of correlated traits were identified from pairwise genetic correlation analysis. These findings have substantial implications for clinical applications. By identifying shared genetic risks across diseases, at-risk individuals can be alerted to pursue early detection of comorbidities and targeted prevention strategies. For example, the clustering of cardiometabolic traits, such as type 2 diabetes, hypertension, and BMI, highlights their interconnected genetic basis and suggests that individuals with a high genetic risk for one condition may benefit from early screening and intervention for related conditions[13]. Additionally, the shared genetic architecture allows the development of multi-trait PRS models that integrate genetic risks across correlated traits, improving prediction accuracy and enabling precision medicine approaches that address multiple health outcomes simultaneously[14]. Including the correlated traits in PRS model development improved performance, resulting in an average 1.54-fold increase in the explained percentage of phenotypic variation.

13. I commend the authors for describing PRS performances in the context of upper limits of SNP-heritability (lines 516-535).

Lines 516-535: (Polygenic Risk Score (PRS) Development) Building on these insights, we developed and validated PRS models that demonstrated strong predictive performance for a wide range of diseases. Although we used five PRS tools, including , LDpred2, Lassosum2, PRS-CS, SBayesR, and MegaPRS (Tables S9-S13), we found that LDpred2 outperformed the others for most traits. Therefore, we took the results of LDpred2 for further comparisons. Out of the PRS

models, AUC values surpassed 0.55 for dichotomized phecodes, while all 24 models for quantitative traits accounted for more than 3% of the phenotypic variance. (Table S9 and Extended Data Fig. 4). The top-performing PRS models included well-known heritable traits such as ankylosing spondylitis (AUC = 0.812±0.016), psoriasis (0.710±0.016), atrial fibrillation (0.702±0.014), prostate cancer (0.696±0.018), systemic lupus erythematosus (0.695±0.015), rheumatoid arthritis (0.649±0.011), type 2 diabetes (0.642±0.005), female breast cancer (0.610±0.010), and hypertension (0.610±0.004). Interestingly, the PRS for hepatitis B also demonstrated genetic predictability (0.655±0.008). Comparing these results with heritability, which serves as the theoretical upper bound of PRS explained variance (r^2), we found that 27 phecodes and 10 quantitative traits explained over 60% of their heritability, including prostate cancer, type 2 diabetes, female breast cancer, HDLC, triglycerides, and red blood cell count (Fig. S2).

Response: We have rewritten the indicated sentences (lines 516-535) as suggested.

Revised sentences (in results, LINE 560-571)

Since SNP-heritability (h^2) represents the upper bound of variance that can be explained by PRS, we examined the proportion of heritability captured by our models (r^2/h^2). Several traits, including prostate cancer (r^2/h^2 = 0.054/0.070), type 2 diabetes (0.066/0.126), and HDL cholesterol (0.136/0.191), reached over 50% of their SNP-heritability, suggesting that PRS can achieve near-optimal predictive accuracy for highly heritable traits. However, for complex diseases influenced by both genetic and environmental factors, PRS performance is inherently constrained by the fraction of heritability attributable to common variants. These findings reinforce the importance of SNP-heritability as a reference point for evaluating PRS utility and highlight the need for larger, ancestrally diverse datasets to further enhance genetic prediction models (Extended Data Fig. 5).

14. line 593: how about other PGS from the catalog that include larger sample sizes over UKB? Is TPMI-derived PRS still better than the most accurate European PRS from PGSCatalog?

Line 592-594: (Discussion) Indeed, for the 128 traits where there is sufficient sample size in the cohort, the PRS performance rival those developed for Europeans using UKB data.

Response:

To address the reviewer's concern regarding the performance of TPMI-derived PRS compared to larger-sample-size European PRS from the PGS Catalog, we conducted additional evaluations and incorporated the results into the revised manuscript.

As shown in Figure R1-6, TPMI-derived PRS generally outperforms or is comparable to the highest-performing PRS models from the PGS Catalog (PGS Catalog-1 and PGS Catalog-2). Specifically, for prostate cancer, type 2 diabetes, and alcoholism, TPMI-derived PRS achieves a higher AUC than both European-derived PRS models. However, for certain traits such as glaucoma and migraine, the performance differences are less pronounced, indicating that PRS generalizability varies across diseases.

These findings reinforce the importance of population-specific PRS training, as models optimized using ancestry-matched cohorts tend to achieve higher predictive accuracy than those derived from larger but non-matched datasets. Since the results are similar to what get from UKB-derived models, we decided to stick with the comparison with UKB for the simplicity and clarity.

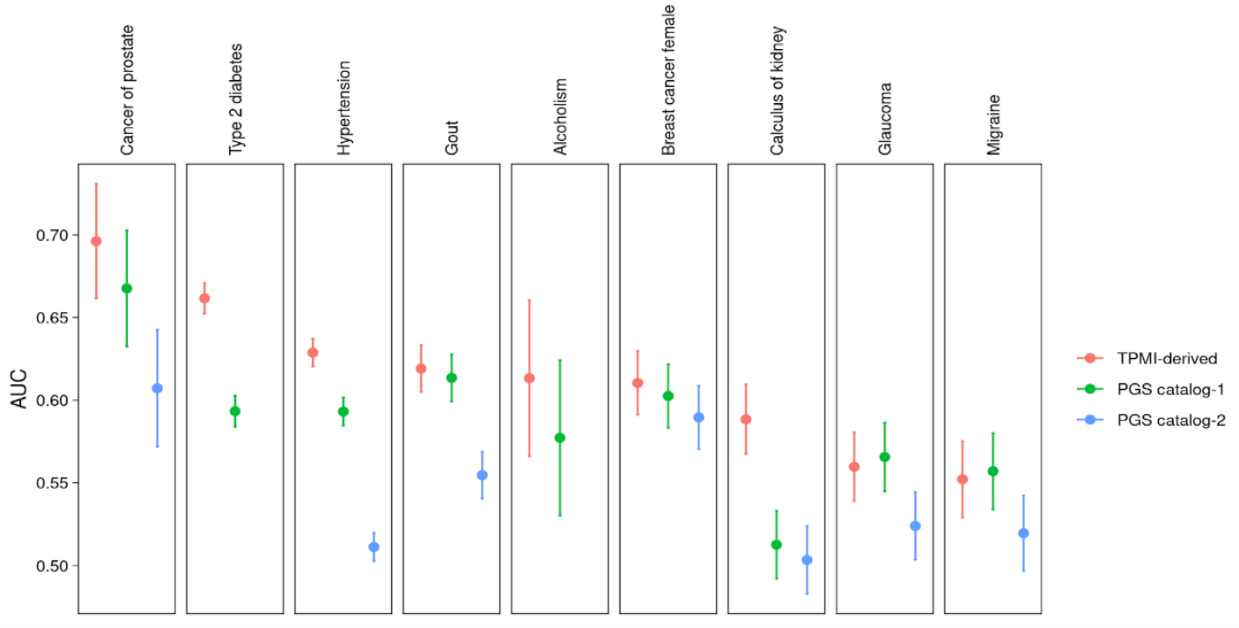


Figure R1-6. The PRS performance of TPMI-derived models and models from PGS catalog.

Table R1-5. The detailed information of PRS models from PGS catalog

			Sample size	
Trait	Label	PGS ID and Ref.	GWAS	Training
Prostate cancer	PGS catalog-1	PGS000590 (Fritsche et al. 2020)[15]	175,928	5,650
Prostate cancer	PGS catalog-2	PGS002241 (Mars et al. 2022)[16]	140,254	175,364
Type 2 diabetes	PGS catalog-1	PGS000330 (Mars et al. 2020)[17]	898,130	21,813
Type 2 diabetes	PGS catalog-2	PGS004859 (Deutsch et al. 2023)[18]	1,114,458	
Hypertension	PGS catalog-1	PGS002765 (Mars et al. 2022)[19]	757,601	
Hypertension	PGS catalog-2	PGS003018 (Ma et al. 2022)[20]	361,194	23,307
Gout	PGS catalog-1	PGS003329 (Sumpter et al. 2023)[21]	332,370	
Gout	PGS catalog-2	PGS004768 (Truong et al. 2024)[22]		37,851
Alcohol use disorder	PGS catalog-1	PGS002738 (Lai et al. 2022)[23]	121,604	
Breast cancer	PGS catalog-1	PGS000514 (Fritsche et al. 2020)[15]	213,515	7,368
Breast cancer	PGS catalog-2	PGS003380 (Namba et al. 2023)[24]	247,173	
Calculus of kidney and ureter	PGS catalog-1	PGS001250 (Tanigawa et al. 2022)[25]	269,704	
Calculus of kidney and ureter	PGS catalog-2	PGS004493 (Jung et al. 2024)[26]	174,489	
Glaucoma	PGS catalog-1	PGS002043 (Prive et al. 2022)[27]		391,124
Glaucoma	PGS catalog-2	PGS002761(Mars et al. 2022)[19]	351,696	
Migraine	PGS catalog-1	PGS004797 (Truong et al. 2024)[22]		37,851
Migraine	PGS catalog-2	PGS004799 (Truong et al. 2024) [22]		37,851

15. lines 673-680: this is too broad of a claim. There is no formal validation in the manuscript for this hypothesis.

Lines 673-680: (Discussion) This study formally validates the common belief that population-specific risk-prediction models perform well in that population, pointing to the need to build these models in the major populations across the world. As the PRSs we developed perform well in EAS, the implementation of genetic findings is now available to an additional 25% of the world's population. It is hopeful that if all EAS obtain their genetic profiles and determine their risk for major diseases, many diseases can be prevented or their onset can be delayed significantly, thereby fulfilling the promise of modern genetics.

Response: While it is unclear what constitutes an acceptable "formal validation" of the common belief that population-specific risk-prediction models perform well in that population; our data definitely points in that direction. In response to the reviewer's suggestion, we have revised the text to narrow the scope of our claim as follows.

Revised sentences (in discussion, LINE 775-794)

This study demonstrates that population-specific risk-prediction models, such as those developed for EAS in this work, can achieve strong predictive performance for traits with high relevance in that population. The PRSs we developed for EAS performed well for multiple traits, including diseases with significant public health implications, such as type 2 diabetes and systemic lupus erythematosus. However, for certain traits, such as glaucoma, PRS derived from both UK Biobank and TPMI performed comparably, suggesting that the genetic architecture of some traits allows for generalizable models. These findings emphasize the importance of developing and validating PRS models in diverse global populations to maximize their utility and equity in genetic risk prediction. While our results emphasize the utility of developing population-specific PRS, further research is needed to directly compare their performance with multi-population models and assess their generalizability and to assess their impact on disease prevention and management. In particular, longitudinal studies and real-world implementations will be critical to determine the extent to which PRS-guided interventions can delay disease onset or improve health outcomes. Furthermore, it is hopeful that if all can obtain their genetic profiles and determine their risk for major diseases, many diseases can be prevented or their onset can be delayed significantly, thereby fulfilling the promise of modern genetics.

16. A chip-GWAS was performed (lines 737-739). Any details on how the results of that scan were used to minimize bias? Was there a "sample collection site"-GWAS performed as well?

(Phasing and Imputation) In addition, we also performed a chip-GWAS for minimizing the bias from different chips, resulting in a dataset of 8,046,864 well-imputed common genetic variants.

Response: To minimize the effect of the two versions of genotyping chips, we conducted a logistic regression with the version of chip as outcome for all imputed variants. In the regression, we also included 10 genetic PCs as covariates. Any genetic variants with p-value < $1x10^{-5}$ was excluded, and 8,046,864 well-imputed common genetic variants were retained in our main analyses. The detailed methods and Manhattan plot have been included in the revised supplemental document (Figure R1-7 [Figure 8 in the supplementary document p.15]). We didn't conduct the "sample

collection site-GWAS", instead we included the genotyping site (by genotyping batch) GWAS in our QC pipeline and the results were presented in our original supplemental document (Figure 8 in the supplemental document). To minimize the bias of sample collection site, we involved the recruiting hospital, which is also the sample collecting site, as covariates in our main analysis.

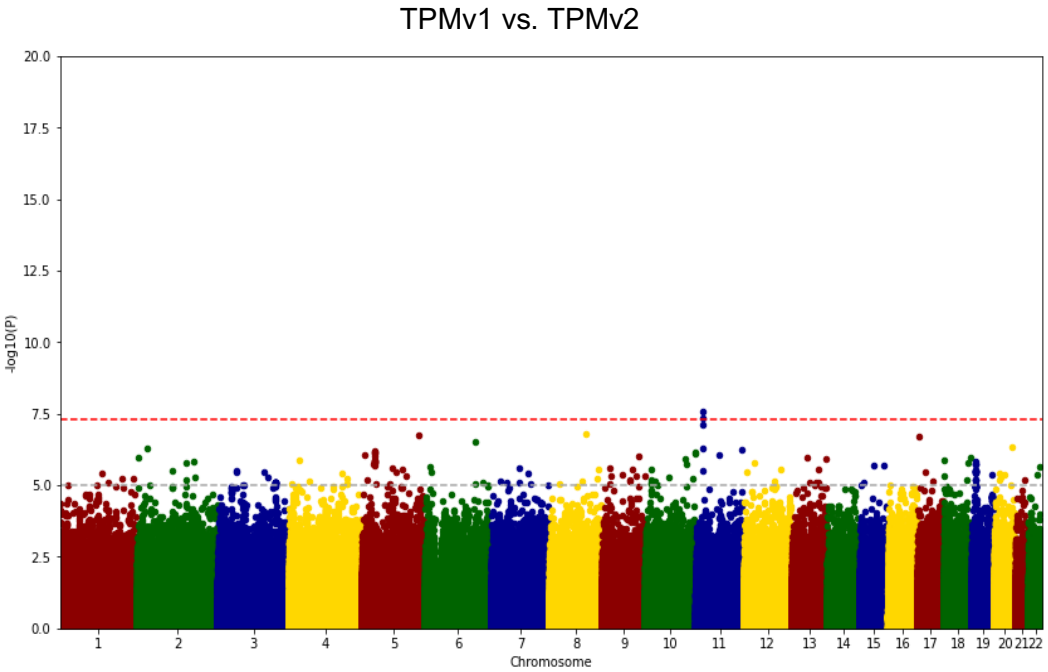


Figure R1-7. (Figure 8 in the supplemental document) The post-imputation Manhattan plot of the comparison between TPMv1 and TPMv2

17. Distribution of relatedness in TPMI is needed.

Response: Per the reviewer's suggestion, we have added a paragraph describing the determination and distribution of genetic relatedness in the supplementary document (page 16).

> To avoid the bias of potential population stratification and remove the close genetic relatedness among TPMI, we leverage the principal component analysis (PCA) and proportion of identity-by-descent (IBD) to capture the population-level genetic structure and close genetic relatives. We firstly identified ancestry representative unrelated subset and conducted PCA among the identified subset for obtaining the robust PC weightings. Then, we predicted the PC scores for all remaining individuals. Following, we estimated the proportion of IBD and kinship with 20 genetic PCs for the adjustment of population structure. The PCA and relatedness estimation was performed with PCAiR and PC-Relate. As a result, we detected 80,699 pairs of 1st degree of relatives ($\hat{\pi} > 0.375$), 82,487 pairs of 2nd degree ($0.375 \geq \hat{\pi} > 0.1875$), and 186,173 pairs of 3rd degree ($0.1875 \geq \hat{\pi} > 0.1$) (Fig 9). In terms of individuals, we have 123,277 individuals (26.6%) have at least one first degree relatives in the cohort, 207,578 individuals (44.8%) have at least one first and/or second degree relatives, and 321,755 individuals (69.4%) have at least one relatives closer then third degree in the cohort.

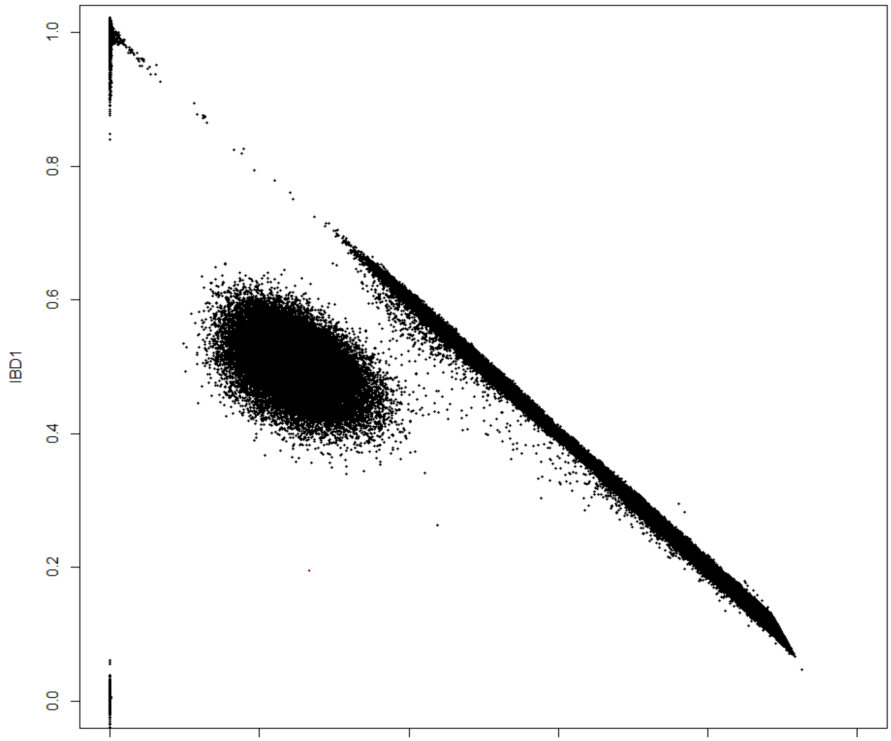


Figure R1-8. (Figure 9 in the supplementary document) The scatter plot of IBD0 versus IBD1 shows pairwise relationships of close and distinct relatedives ($\widehat{\pi} > 0.05$) in TPMI

18. Unclear how PRS was trained. Why was only 100k of TPMI used for PRS instead of the whole sample of 500k? or at least the 250k that are unrelated? (lines 839-841)

Lines 839-841: (Single and Multi-Trait Polygenic Risk Score (PRS)) The preserved dataset of 100,000 unrelated TPMI subjects was split into two subsets, training (n = 80,000) and validation (n = 20,000) for PRS model building.

Response: To ensure robust and unbiased PRS performance, we followed established practices for sample partitioning, model training, and cross-population evaluation[8-10].

Rather than excluding 100,000 unrelated individuals, we allocated them to independent training (N = 80,000) and validation (N = 20,000) sets instead of using them in the GWAS discovery cohort (N = 363,447). This design prevents data leakage, avoids overfitting, and ensures generalizability. Maintaining separate datasets for discovery, training, and validation strengthens the reliability of PRS, as discussed in our response to Question #4, where we emphasized PRS as a complementary tool rather than a standalone predictor.

Revised sentences (in methods, LINE 866-881)

The entire dataset was divided into three subgroups: the GWAS set (N = 363,447), the training set (N = 80,000), and the testing set (N = 20,000) …... Within the GWAS set, we selected an unrelated subset (N = 248,754) to perform GWAS using a generalized linear model in PLINK2, and we conducted 1:10 age, sex-matching for the traits with imbalanced case/control ratio (<1/20). These GWAS statistics were then used for heritability and genetic correlation estimation.

19. lines 863-883: it is unclear what are the sample sizes of the cohorts used in this study. How many participants from All of US where used for this study? Why was a PRS trained on 80k samples in TPMI compared with a sample size of 120k from TWB? How about a PRS that uses all samples together; wouldn't those be the best PRS for this population?

Lines 863-883: (External Validation and Comparison) We conducted an external validation of our developed PRS using data from the Taiwan Biobank, EAS from UK Biobank and All of Us. TWB is a community-based biobank, and it has recruited over 200,000 participants in Taiwan. Herein, we used 120,460 independent subjects, who were genotyped with the Axiom customized chip TWB2 (equivalent to TPMv1), and their genotyping QC, phasing, and imputation followed the same protocol as described above. The self-reported disease condition was queried from their baseline questionnaire, except for cancer. Since the study design of TWB excluded cancer patients at recruitment, we used both baseline and follow-up self-reporting data to define cancer cases and controls. UKB has enrolled ~500,000 participants since 2006 and linked their genetic data with enriched phenotypic data. For external validation, we only used self-reported Chinese (more diverse than Han Chinese) as East Asian and their inpatient record for case definition. All of Us intends to enroll over 1 million participants in the United States and has released whole genome genotyping data for ~312,000 participants as of the first quarter of 2024. The genetically confirmed EAS as well as other super populations and their linked EMR were used for validating our PRS models. Moreover, we compared the TPMI-derived PRS model with UKB-derived models to investigate the performance of population-specific PRS. The UKB-derived models were based on published UKB European GWAS (https://pheweb.org/UKB-TOPMed/), and LDpred2-auto was applied for model building

Response: To address these concerns, we have clarified the Methods Section by specifying the sample sizes used for each dataset, explaining the rationale for training PRS on 80,000 TPMI samples while validating with 20,000 TPMI unrelated subjects and 118,503 TWB subjects, and justifying why we did not integrate all datasets into a single PRS model. Specifically, we now state that 6,895 genetically inferred East Asians from All of Us and 1,527 from UK Biobank were included for external validation, ensuring transparency in cohort selection and consistency in quality control procedures.

The choice to train PRS on 80,000 TPMI samples while validating with 120,000 TWB samples was made to balance computational feasibility and model optimization while maintaining a fully independent validation set. Although TWB contained a larger sample size in this analysis, we ensured that TPMI's PRS was trained using a separate, well-powered sample to facilitate a rigorous and unbiased comparison across datasets. This approach prevents overfitting and ensures that PRS performance is evaluated independently. While integrating all datasets for PRS model training could theoretically enhance predictive power, differences in genotyping platforms, imputation methods, population structures, and phenotypes definition introduce systematic biases that may compromise model reliability. Moreover, the development of medical phenome PRS models in TPMI, which can be obtained from linked EMR, would be limited by the number of disease querying in the TWB questionnaire.

The following is now replaced the original line 863-883 with specific sample size of UKB and All of Us.

Revised sentences (in methods, LINE 1005-1020)

UKB has enrolled ~500,000 participants since 2006 and linked their genetic data with enriched phenotypic data. For UKB validation, we used their inpatient record for case definition. Their ancestral population was determined by self-reported ethnic background, such as self-reported Chinese as EAS (n=1,572), White, British, Irish, and any other white background as EUR (n=472,869), Black or Black British, Caribbean, African, and any other Black background as AFR (n=8,074), and Asian or Asian British, Indian, Pakistani, Bangladeshi, and any other Asian background as SAS (n=9,893). All of Us intends to enroll over 1 million participants in the United States and has released whole genome genotyping data for ~312,000 participants as of the first quarter of 2024. We applied ADMIXTURE with 1000 Genomes as reference panel for assigning the genetically inferred ancestral populations, including EAS (n= 6,895), EUR (n= 152,754), AFR (n= 60,964), AMR (n= 32,394), and SAS (n=2,334). The genetically confirmed EAS as well as other superpopulations and their linked EMR were used for validating our PRS models.

Referee #2 (Remarks to the Author):

The authors of the manuscript "Population-Specific Polygenic Risk Scores Developed for the Han Chinese" have done a tremendous amount of work in the Taiwan Precision Medicine Initiative. This is an important resource and I believe the manuscript will be a worthy contribution to the field, showing how diverse biobanks can bring both biological insights and necessary tool development for precision medicine. Overall, the paper would be greatly improved if the authors took up some more real estate going through the specific examples as to why novel discoveries were made (such as is it related to allele frequency differences or different epidemiology of that specific trait), as well as clarifying some vague statements. Right now there is a lot of work that has been included, but almost everything is up to the reader to do a deep dive of every gene without some exemplars demonstrating how important this work is. I have separated out my comments below into major and minor concerns. Some things I have noted as not necessary to the paper, but would be a nice addition.

Response: We thank the reviewer for careful reading of our manuscript and recognizing the value of our work. We appreciate the insightful suggestions and concerns raised. We have taken the comments to heart and have addressed them as detailed below. We have revised the manuscripts per the suggestions as appropriate.

1. The manuscript would be substantially strengthened by delving into the reasons why there are novel signals and relationships within the TPMI compared to much larger non-Han Chinese biobanks. This could take the form of allele frequency and/or effect size comparisons, or comparison of prevalence of different outcomes between the biobanks. While the included analyses are substantial and a very worthy contribution to the field, going a bit deeper and trying to explain it would be a huge contribution and make a better argument for the need for diverse biobanks. This would be necessary throughout the manuscript. Right now a lot of it is focused on discovery, which again while this would create a resource of known associations, if the central premise of the paper as per the abstract is that diverse biobanks are needed, more real estate should go into explaining why that is beyond the yield of novel associations. Overall, the paper would be greatly strengthened by putting their results into context of the field at large, such as pop-specific vs multi-pop PRS or what is known about the genetics of specific outcomes (such as HBV).

   Response: Per the helpful suggestions, we have made the changes following the reviewer's guidance. We firstly examined the allele frequencies of novel SNP in TPMI and other populations, then tested ancestry-specific effect size, and following the disease prevalence between different cohorts.

   Revised sentences (In results, LINE 450-480)

   Of the 95 novel genetic associations, 30 variants are rarely found (MAF<0.05) in other populations (African, Admixed American, South Asian, and European (non-Finnish) in gnomAD), and 33 variants less than 0.01 in the Europeans, the most extensively studied population, which explains why they were not reported in previous GWAS. For example, the SNP rs17089782, a missense variant in the *PIBF1* (p.R405Q) gene on chromosome

13 is significantly associated with thyroid cancer ($p = 2.8x10^{-9}$) in the TPMI cohort. This SNP, rs17089782, has a minor allele frequency of 5.65% in TPMI but 0.01% in European populations. This population-specific difference likely explains why this association was detectable in TPMI but not in predominantly European datasets. However, *PIBF1* is essential for immune regulation, especially during pregnancy, and is relevant to autoimmune diseases and cancer[5]. Another novel variant identified in our quantitative analysis of BMI (rs761018157 p-value = $4.8x10^{-9}$, MAF in TPMI = 4.34%, MAF in EUR<0.01%,) maps to *PHOX2B*. This gene, highly expressed in the nervous system, had previously been linked to obesity hypoventilation syndrome in a small study (n = 30)[6] and associated with bone mineral density[7]. In addition, when we compare the effect size in TPMI and UKB for the rest of the novel findings, 25 exhibit a significant different effect size (p-value<0.05). For instance, a TPMI-identified platelet count-associated variant, rs12955741, located in the intergenic region between *TGIF1* and *DLGAP1* exhibits a different effect size compared to that in UKB (β_{TPMI} = 0.044, β_{UKB} = -0.005, p-value = $1.7x10^{-9}$). Moreover, the high hepatitis B virus (HBV) carrier rate in Taiwan[28] contrasts sharply with its rarity in European cohorts, enabling TPMI to identify novel loci associated with viral hepatitis B (case number in TPMI = 23,618 vs. UKB = 132). Among the 26 independent loci identified in our analysis of hepatitis B, 19 fine-mapped loci are novel (Extended Data Fig. 2). Notably, 18 of these 19 loci were found to be associated with liver function or diseases (Table S5). These novel associations highlight the uniqueness of certain disease loci in the TPMI cohort, presenting opportunities for developing population-specific therapeutic interventions and advancing precision medicine.

2. Was the correlation between the National Health Insurance Research Database different by trait category? This may identify some bias in terms of the recruitment of the 16 medical centers. Also, were there differences by sites? (This last part is not necessary for the manuscript, I just think it would be an interesting piece tied to why TPMI is different or as a citation for future TPMI papers.)

   Response: We appreciate the reviewer's inquiry regarding whether the correlation between TPMI and NHIRD disease prevalence varies by trait category, as this may provide insights into potential ascertainment biases associated with TPMI's recruitment strategy. To address this concern, we conducted additional analyses and examined the correlation between TPMI and NHIRD disease prevalence across different trait categories (Table R2-1). Our results indicate substantial variation in the following table and these findings suggest that certain disease categories, such as neoplasms (r = 0.816), exhibit stronger alignment with NHIRD prevalence, whereas others, such as hematopoietic disorders (r = 0.618) and mental disorders (r = 0.656), demonstrate weaker correlations.

   We acknowledge that the correlation of 0.658 between TPMI and NHIRD disease prevalence indicates a moderate alignment rather than full population representativeness. To ensure a more precise interpretation of this result, we have explicitly discussed ascertainment biases and demographic differences in the revised manuscript.

To further explore potential ascertainment biases, we analyzed the demographic distributions of TPMI participants and compared them to NHIRD. We found that TPMI participants are overrepresented in the middle-aged group (year of birth 1941-1970), comprising 54.3% of the cohort, compared to 38.3% in the NHIRD. Additionally, TPMI includes a slightly higher proportion of female participants (55.1%) compared to NHIRD (50.6%). (Figure R2-1) Furthermore, TPMI has a higher proportion of participants from northern Taiwan, with 59.5% recruited from hospitals located in northern Taiwan, compared to 47.3% in the NHIRD from individuals residing in northern Taiwan. In contrast, less TPMI participants are from southern Taiwan (10.3% vs. 27.4%), These demographic differences, combined with TPMI's volunteer-based recruitment process, likely contribute to the lower proportion of cases observed in TPMI compared to NHIRD.

Table R2-1. The correlation coefficients of each phecode category

Category	Number of phecodes	Correlation coefficient (r)	p-value
Infectious diseases	24	0.716	8.26E-05
Neoplasms	74	0.817	7.65E-19
Endocrine/metabolic	55	0.720	6.02E-10
Hematopoietic	15	0.618	1.40E-02
Mental disorders	33	0.656	3.41E-05
Neurological	28	0.655	1.53E-04
Sense organs	58	0.651	3.18E-08
Circulatory system	62	0.806	2.60E-15
Respiratory	43	0.596	2.51E-05
Digestive	76	0.801	1.24E-13
Genitourinary	82	0.588	6.30E-09
Dermatologic	39	0.640	1.15E-05
Musculoskeletal	50	0.775	4.04E-11
Congenital anomalies	5	0.867	5.73E-02
Symptoms	22	0.593	3.61E-03
Injuries & poisonings	29	0.493	6.63E-03
Total	695	0.656	2.69E-84

Revised sentences (In discussion, LINE 722-746)

As other large-scale epidemiological studies, ascertainment bias is also observed in our study. TPMI's case proportion also shows a moderate but significant correlation with the prevalence from National Health Insurance dataset, and this may imply the ascertainment bias of TPMI's hospital-based design. Compared to the general population in Taiwan (NHIRD), TPMI participants are overrepresented in the middle-aged group (year of birth

1940-1970, 54.3% vs. 38.3%), include slightly more females (55.1% vs. 50.6%), and have a higher proportion of participants from northern Taiwan (59.5% vs. 47.3%). These demographic differences, along with the volunteer-based recruitment process, likely contribute to the lower case proportions observed in TPMI (Figure 1a). Importantly, a significant portion of disease records in the NHIRD originate from local clinics and primary care settings, which are not covered in TPMI. Notably, the EMR of the participants are incomplete, as some participants receive care from multiple health providers, but the TPMI only has access to EMRs from their enrollment hospitals. We acknowledge that ascertainment biases may influence disease prevalence and heritability estimates. Thus, we accounted for case-control ascertainment in heritability estimates and PRS evaluations by applying liability-scale transformations using population prevalence data from NHIRD and utilized independent validation datasets for PRS validation (TWB, UKB, All of Us) to mitigate the impact of ascertainment biases, while acknowledging that residual biases may persist. Methods like inverse probability weighting could mitigate such biases[4], but these require detailed external reference data that are currently unavailable. This limitation also emphasizes the need for future adjustments to enhance generalizability.

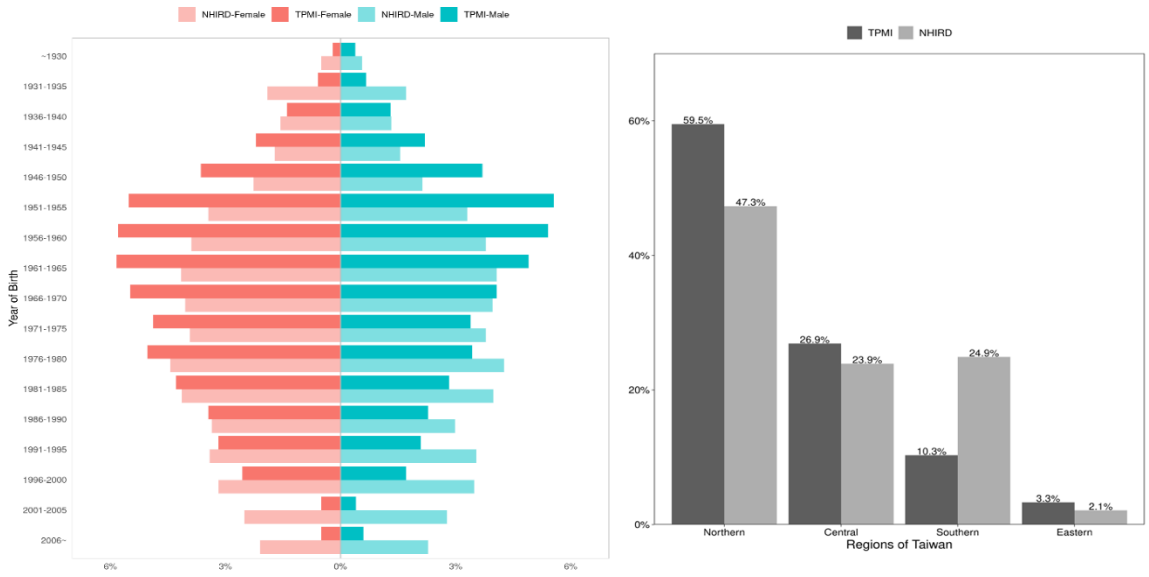


Figure R2-1. The basic demographic information comparison between TPMI and NHIRD for year of birth by sex (A) and regions (B)

3. The manuscript would be strengthened in terms of readability if some of the technical details were integrated into the main text. For example, In line 434 the authors talk about replication rate with other EAS GWAS, but there is no mention of where these GWAS come from, nor is it in the Methods. It took me looking at Extended Data Fig. 1 to find that they are pulled from the GWAS Catalog. How were these selected/filtered from the overall database?

Line 434. we observed high replication rate of reported disease loci from EAS GWAS (74.4%), particularly for endocrine/metabolic and hematopoietic diseases (88.68% and 84.62%, respectively; Extended Data Fig. 1 and Table S3).

Response: For the particular question regarding the source of reported associations for examining the replication rate, we used a published tool, PGRM, to determine the replication rate of our GWAS. PGRM summarizes a list of reported GWAS signals corresponding to phecodes, along with the source of ancestry population based on GWAS Catalog data downloaded on January 4, 2022. The detailed inclusion and exclusion criteria are described in their paper[29]. In brief, they excluded: 1) unqualified phenotypes (e.g., subtypes of diseases, age of onset, or family history); 2) specialized cohorts (e.g., specific patients, mutation carriers, or individuals with special exposures); and 3) studies based on non-standard regression models (e.g., interaction terms or family-based models). As a result, they included 5,879 genetic associations from 523 GWAS publications.

We have revised text to include as many technical details as possible, for example:

Revised sentences (In results, LINE 407-414)

We examined 695 dichotomized phenotypes (phecode, case n > 2,000) and 24 quantitative traits (sample size > 100,000), spanning numerous disease categories (defined by phecode groupings[30,31]), such as neoplasms, metabolic disorders, circulatory conditions, autoimmune diseases, and more (Fig. 1). The phecodes, derived from International Classification of Diseases (ICD) codes[30,31], alongside quantitative traits such as blood pressure, BMI, liver enzymes, and lipid levels, provide a robust dataset for exploring genetic contributions to human health (Table S1 and S2).

And LINE 426-430

We observed a high replication rate of reported disease loci from EAS GWAS on GWAS catalog (actual/expected ratio (AER) = 78.17%, considering the statistical power with the published tool, PGRM[21]), particularly for endocrine and metabolic/hematopoietic diseases (AER = 88.68% and 84.62%, respectively; Extended Data Fig. 1 and Table S3).

4. Another example is the use of "independent" without saying what independent actually means (R2<X, etc). While some of these details are found in the Methods, there's no harm (and would actually help the reader) to have some of them in the main results text.

Response: To clarify this suggestion, we report the independent causal variant which are identified by SuSiE fine-mapping. We reported the genetic variant with highest posterior inclusion probability (PIP) for the 95% credible sets, and the single lead variant is reported for the failed fine-mapping regions and MHC region. Also, we apologize for a bug in our counting script, which resulted in an undercount of the lead variant in the failed fine-mapping region of the GWAS with successful fine-mapping regions. After correcting this error, we now report 1,309 variants from the phecode GWAS and 1,347 variants from quantitative traits. Per the reviewer's suggestion to clarify a key term, we have added a brief explanation of fine-mapping independence criteria.

Revised sentences (in results, LINE 434-446)

We applied the sum of single effects model for fine-mapping, and the reported the genetic variant with highest posterior inclusion probability of identified credible sets and the single lead variant for the failed fine-mapping regions and MHC region. Our analyses revealed a total number of 2,656 fine-mapping identified independent association signals, including 1,309 from phecodes GWAS and 1,347 signals from quantitative traits. …… In addition, we identified 217 new hits from previously reported regions, defined as having r2<0.1 with any variant observed in the NHGRI-EBI GWAS Catalog within 1 MB for the same phenotype (Table S4 and S5).

5. The investigators use the threshold of 5X10E08 for genome-wide significance, however there are >700 GWAS being conducted and therefore this does not account for multiple comparisons across analyses. Would the changing of this threshold change the conclusions at all? If not, then I would just say that. If so, how?

   Response: To address potential concerns about multiple comparisons, we conducted sensitivity analyses using a Bonferroni-adjusted threshold ($p < 6.95 \times 10^{-11}$). Our results remain consistent, with the majority of significant loci remaining robust under this stricter threshold. We will briefly mention this robustness in the Results section.

   Revised sentences (in results, LINE 446-449)

   After applying multiple test correction, 1,502 fine-mapped associations passed Bonferroni-adjusted threshold ($5x10^{-8}$/ (695+24) = $6.95x10^{-11}$), as well as 21 novel variants and 115 new hits.

6. Some of the heritability estimates seem substantially lower than other biobanks/studies, such as height which has a heritability of 32.3%. What is the interpretation of this?

   Response: Regarding the heritability estimates for height and other traits, we compared our heritability estimates for the 24 quantitative traits with those from Taiwan Biobank, Biobank Japan, and UK Biobank[12], which used the same statistical approach (Figure R2-2). The results suggest that our estimates are consistent with previously reported heritability. Although the heritability of body height is slightly lower than the estimate

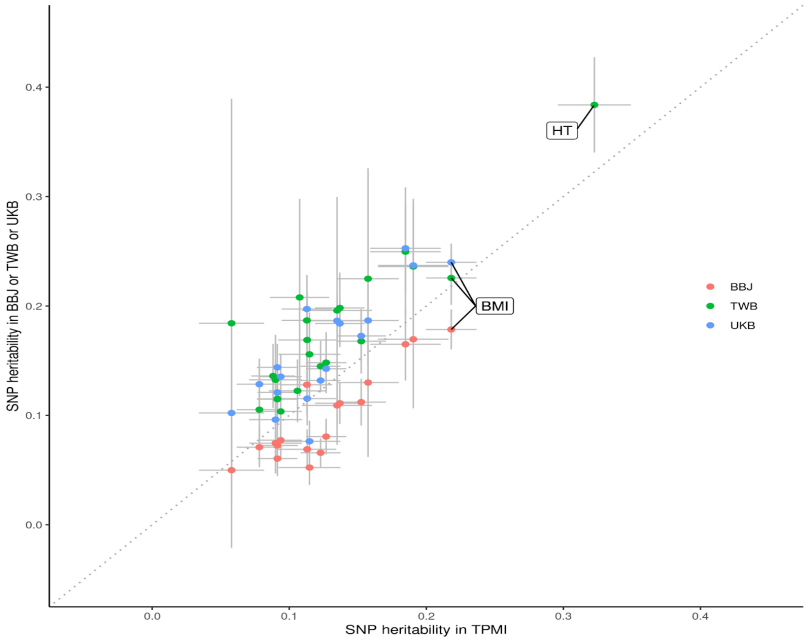


Figure R2-2. The estimates of heritability for 24 quantitative traits in TPMI (x-axis) and other biobanks (y-axis), including Taiwan Biobank (TWB, green), Biobank Japan (red), and UK Biobank (blue).

from the Taiwan Biobank, this may be explained by the fact that the EMR-extracted body height includes both self-reported and measured values, which may introduce more phenotypic variation compared to the Taiwan Biobank, where measurements were taken by a team of trained researchers following the same protocol.

7. I would love a bit more explanation of how to interpret/interesting findings from Figure 3 with an expansion of the results in the main text. Right now it's a bit of a data dump, which while understandable makes it hard to parse what readers should find super interesting.

   Response:

   To address the reviewer's feedback, we have expanded the manuscript to provide a clearer interpretation of the key findings in Figure 3. Specifically, we emphasize the functional relevance of gene-level heritability partitioning and colocalization analysis, highlighting genes with pleiotropic effects across multiple traits and their potential therapeutic implications.

   Revised sentences (in discussion, LINE 498-520)

   We then partitioned the heritability at the gene level and identified 329 unique genes contributing significantly to phenotypic variation (h^2 > 0.1% and z-score > 1.64). Of these, 45 affected more than one phecode and/or quantitative category, including key genes such as *APOE, APOC1, TOMM40, ABCG2,* and *KCNQ1* (Fig. 3 and Table S7). We also conducted a colocalization analysis to elucidate the potential molecular function of identified GWAS signals with three QTL datasets, including GTEx[32], MAGE[33], and JCTF[34]. (Fig. 3 and Table S8). Our results identified 391 unique genes that significantly mediate the outcome through their expression (posterior probability > 0.9), including *GBAP1* which colocalized with five different traits (uric acid, serum creatinine, hematocrit, hypertension, and gout). Among the colocalized genes, 75 of them can be identified only in the multi-ancestry lymphoblastoid cell lines eQTL, MAGE (20 genes), and/or Japanese whole blood eQTL, JCTF (59 genes). Our findings demonstrate the effect of these genes (such as *APOE, ABCG2*, and *KCNQ1*) on multiple traits and disorders (Table S6). For example, *APOE* influences both lipid metabolism and Alzheimer's disease risk, making it a promising target for therapeutic strategies aimed at managing these interconnected conditions. By elucidating shared genetic effects, these results offer opportunities to develop precision medicine approaches that address comorbidities, such as treating hyperlipidemia and reducing dementia risk through a single intervention targeting *APOE.* This gene-level understanding emphasizes the potential to optimize therapeutic strategies by leveraging genetic pleiotropy in disease management.

8. 1) The use of phecodes that are highly correlated and even could be nested (such as psoriasis, psoriasis and related disorders, psoriasis vulgaris, psoriasis athropathy) is a bit misleading when it comes to the genetic correlation and clustering. With so much overlap in cases, of course they will genetically overlap. For these analyses I could see two possible avenues. One would be to remove highly correlated phecodes (such as T2D and 7 different T2D w/specific manifestations/symptoms) to get a better idea of how these clusters inform each other, OR to focus on what is different or specific to these phecodes.

For example, are there differences in genes associated with T2D w/peripheral circulatory disorders or T2D with renal manifestations?

2) The above concern re:redundant phecodes may be particularly concerning with the PRS training using multiple phecode-specific PRS. As I interpret the manuscript currently, the authors only compared the phecode-specific PGS developed in TPMI with external validation in the other large-scale Taiwan study and UKB/AoU. PRSmix has the possibility of overfitting which would be demonstrated with this external validation.

Response: We fully understand the reviewer's concern regarding the correlation between phecodess, so we reran the genetic correlation estimation and clustering after removing redundant phecodes. The results, shown in the following plot (Figure R2-3), indicate that the three main clusters are still observed.

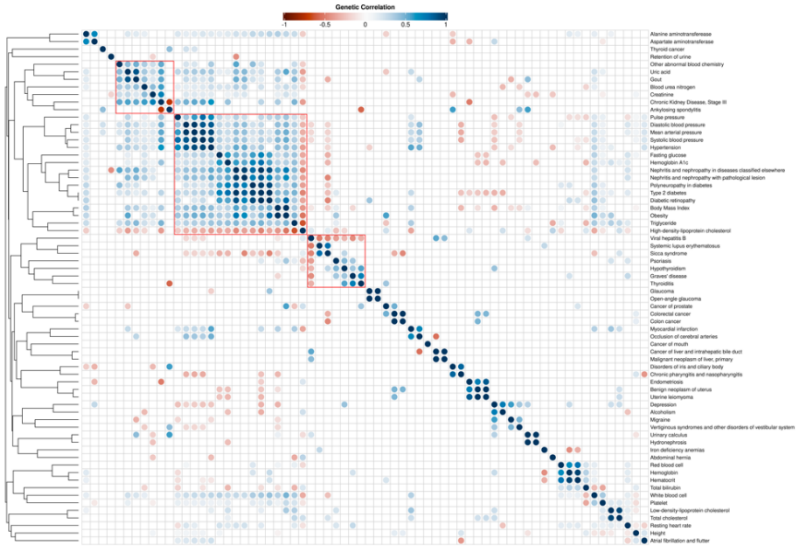


Figure R2-3. The heatmap of pair-wise genetic correlation after removing redundant phecodes, and the clustering tree indicating the distance based on genetic correlation.

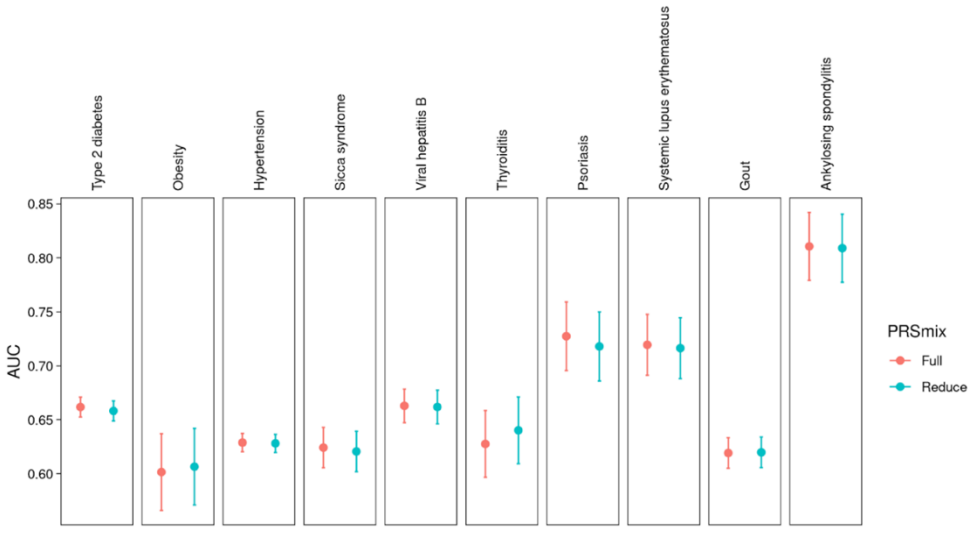


Figure R2-4. The performance of PRS from PRSmix with full pheodes (red) and without redundant phecodes (green)

Next, we conducted a PRSmix+ analysis based on these clusters and found that the performance of PRS models was almost identical when compared to our original models. Since no significant harm from redundant phecodes was observed, we have decided to report our results using the original phecode system, as it is easier to compare with other studies that also apply phecodes. (Figure R2-4)

9. For the PGS methods, please add the actual numbers for the AUC instead of relative improvements (see lines 541-542). Relative improvement is not as relevant to interpretation as the actual AUC.

   Lines 541-542: (Polygenic Risk Score (PRS) Development) The performances of autoimmune and kidney-related clusters were also enhanced, with AUC improvements of 2.1% and 0.8%, respectively, and 1.44- and 1.32-fold improvements in variance explained (r^2).

   Response: Per the suggestion of including absolute AUC values for better interpretation of PRS performance, we have revised lines 541–542 to report the actual AUCs alongside their improvements.

   Revised sentences (in result, LINE 572-580)

   Leveraging the identified clusters, we performed a multi-trait PRS training, PRSmix+[22], for the traits within each cluster (Fig 5 and Table S14). Notably, multi-trait PRS models improved prediction accuracy for the cardiometabolic disease cluster with a 0.040 increase in AUC (from 0.608 to 0.648) and a 1.770-fold improvement in phenotypic variance explained (r^2). The performances of autoimmune and kidney-related disease clusters were also enhanced, with averaged AUC improvements of 0.018 and 0.009, respectively (from 0.641 to 0.659, and 0.601 to 0.610 respectively), and 1.351- and 1.349-fold improvements in variance explained (r^2).

10. The language around external validation and comparison of PRS models is both vague and misleading. For example, lines 561-562 state that the "TPMI-derived PRS models also consistently outperformed UKB European-derived models when applied to EAS", which is shown in Figure 6. However, the confidence intervals for the TPMI-derived and UKB-derived almost always overlap for most of the traits shown in Figure 6. The only exceptions seem to be T2D (although only for the two Taiwan studies), gout (again only for Taiwan-based studies), and migraine (only TWB?). It is fine to say that they performed just as well, or seem to perform better *although the confidence intervals overlap*.

    Response: Per the helpful suggestion, we have revised our statement.

    Revised sentences (in result, LINE 604-609)

    TPMI-derived PRS models perform better than the UKB European-derived models when applied to EAS for viral hepatitis B, type 2 diabetes, hypertension, gout and migraine. For the other traits, TPMI derived models consistently outperform than the UKB ones although the confidence intervals overlap. However, the overlapping confidence intervals in UKB and All of Us may be due to their limited sample size of EAS.

11. Are the PGS adjusted for any other variables, such as age, sex, and study? If so, are the R2 values the incremental R2 of just the PGS or are they the overall performance of the model including these variables? These two have very different interpretations and I didn't see those details in the methods.

    Response: In the original manuscript we only reported the raw R2 without adjusting for any covariates. We have added the description to clarify it in our revised methods. Meanwhile, we also added the age-, sex-included R2 as well as partitioned R2 adjusting for age and sex in the supplemental table. All their definition and detailed method are included in revised methods.

    Revised sentences (in method, LINE 974-985)

    The explained variance (r^2) was used to evaluate the performance of PRS for quantitative traits[35,36], and two indices, area under the receiver operating characteristic curve (AUC) and liability-scaled $r^{2,}$ were used for PRS of dichotomized phenotypes. We followed the approach by Ni et al.[36] and report both raw r^2 and r^2 adjusted for covariates (sex, age, and PCs). Additionally, we include partial r^2 estimates, calculated using the R package rsq. To account for population stratification in cross-cohort predictions, we also report r^2 with PCs as covariates in the supplemental tables. For AUC comparisons, we include a baseline model incorporating standard covariates (sex, age, and PCs) to better assess the added predictive power of PRS. The likelihood ratio test was used to obtain the significance for r^2 with R package, lmtest, and standard error for AUC was calculated with R package, auctestr.

12. The "Impact of Genetic Risks on Overall Disease Burden" is similarly vague and overstating the results through omission. For example, lines 579-580 state that the cardiometabolic cluster contributes to 1.14% of clinical visits and 3.22% of hospitalizations. What is omitted is that is only comparing the top 5% to the bottom 5% and that the values decreased substantially when looking at the top vs bottom 20%. This distinction is important as it isn't that the overall PRS accounts for that proportion looking at linear relationships from both factors, but only the extremes of the PGS. I don't think this weakens the authors points at all! It just needs a bit of explanation to put into context. Also, it would be interesting to know if these results differ by study site to identify selection bias or overall differences between the 16 medical centers.

    Response: Per the reviewer's helpful suggestion, we have revised the relevant section for clarity. In Table 1, we present both the raw explained variance (r^2) without covariates and the partial r^2 after adjusting for sex, age, and recruitment hospital. This helps distinguish the independent contribution of genetic risk from potential confounding factors. Across most conditions, the partial r^2 remains stable or increases after adjustment, reinforcing the robustness of our findings. Also, we provide the partial R^2 of covariates in Table S15 (Table R2-2). The recruiting hospital shows a huge proportion on clinical visit frequency comparing to hospitalization duration.

    Revised sentences (in results, LINE 623-632)

We found that 131 of the top performing PRS models (LDpred2 models with AUC > 0.55 and all PRSmix+ models for phecodes and all models for quantitative traits) are significantly associated with overall health indices, explaining 8.47% of the variation in clinical visit frequency (p-value = 2.69×10^{-14}) and 10.29% of the variation in hospitalization duration (p-value = 5.62×10^{-27}, Table 1 and S15) in the comparison between top and bottom 5% groups after adjusting sex, age and recruiting hospital. Among the identified clusters, the cardiometabolic disease cluster contributed the most to the indices, accounting for 1.32% of clinical visits (p-value = 0.02) and 3.55% of hospitalizations (p-value = 7.10×10^{-9}).

Table R2-2. (Table S15) The partial explained variance (R^2) of covariates in the health indices models.

		Top 5% vs Bottom 5%		Top 10% vs Bottom 10%		Top 20% vs Bottom 20%	
Outcome	Covariate	R^2	p-value	R^2	p-value	R^2	p-value
Clinical Visit	Age	0.138	2.63E-49	0.129	1.59E-90	0.110	1.22E-151
	Sex	0.000	4.56E-01	0.000	9.02E-01	0.000	6.85E-01
	Hospital	0.438	1.45E-165	0.397	7.91E-299	0.336	<1E-300
Hospitalization	Age	0.000	4.77E-01	0.000	4.84E-01	0.000	4.66E-01
	Sex	0.000	9.00E-01	0.000	5.63E-01	0.000	7.57E-01
	Hospital	0.002	9.65E-01	0.001	9.69E-01	0.002	3.05E-01

13. Were any sensitivity analyses conducted such as matching on age or using a longitudinal design on age to account for the 4th limitation in lines 663-665? That may address some of the issues with younger participants not having time to develop the outcome at all? Or any inclusion/exclusion criteria for age?

Lines 663-335: (Discussion) Fourth, some of the younger participants have high-risk genetic profiles but are disease free for those diseases. The duration of the project is too short to determine whether they will eventually develop those diseases.

Response: To address the concern that younger participants may not have had sufficient time to develop late-onset diseases, we conducted age-restricted analyses, limiting the controls to participants aged > 50 years for selected late-onset diseases, including Alzheimer's disease, prostate cancer, and senile cataract.

However, the results from Figure R2-5 showed no significant improvement in risk prediction compared to the full dataset. This suggests that our current modeling approach, which already includes age, age², and their interactions with sex, adequately accounts for age-related effects. These findings imply that more precise phenotyping, rather than simple age restrictions, may be necessary to better capture genetic risk for late-onset conditions.

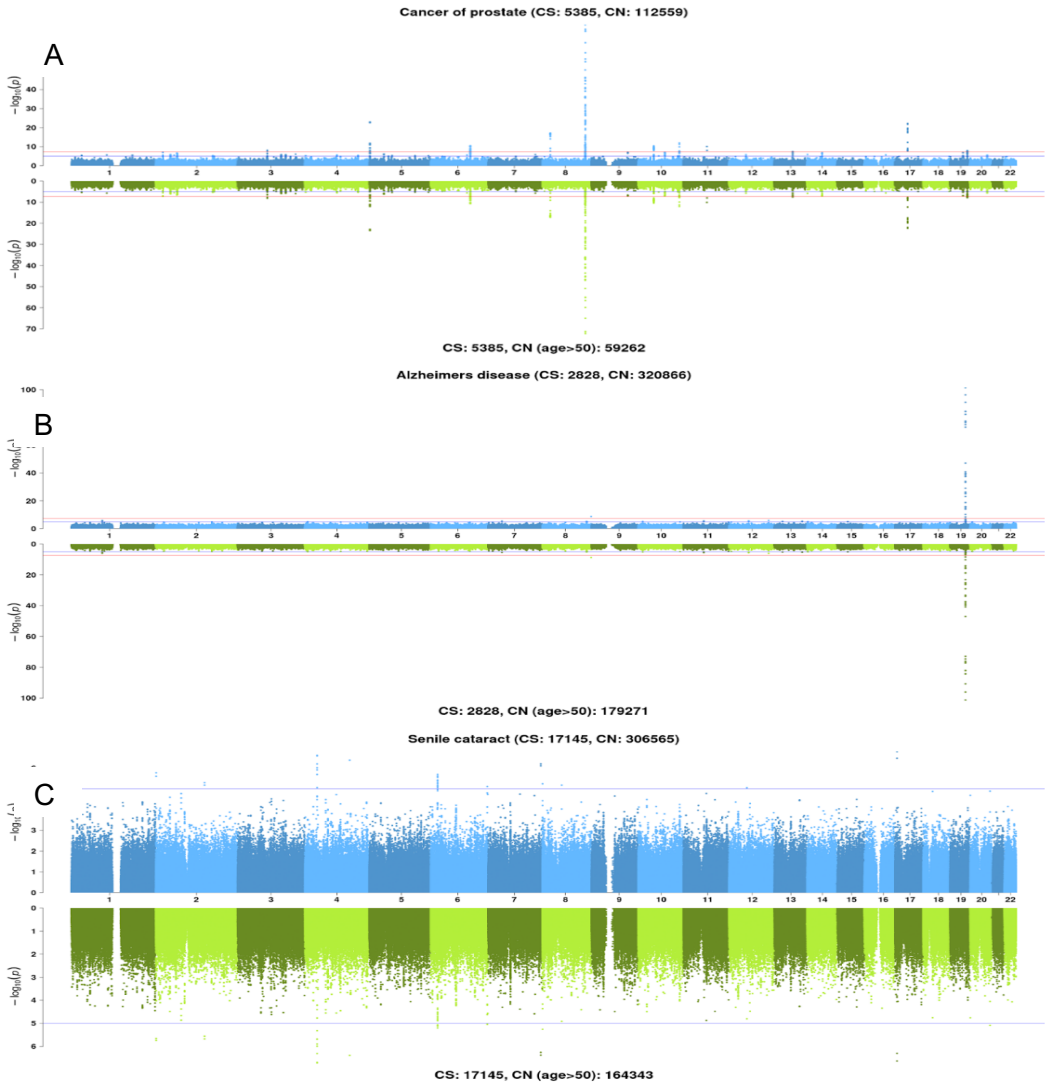


Figure R2-5. The Miami plot for the GWAS of three late-onset diseases without (top) and with (bottom) age restriction for controls: cancer of prostrate (A), Alzheimer's disease (B), and senile cataract (C).

14. The authors state in lines 673 that the study validates the common belief that population-specific risk-prediction models perform well. However, they did not compare their pop-specific PRS to a multi-pop PRS, so cannot make the claim that this means we need pop-specific methods. This is also somewhat counteracted with the paragraph directly after that shows that sometimes the population doesn't matter (equal performance if UKB or TPMI-derived). These two paragraphs could be combined and put into better context with what is known in the field for the development of PGS for these outcomes.

Lines 673: (Discussion) This study formally validates the common belief that population-specific risk-prediction models perform well in that population, pointing to the need to build these models in the major populations across the world.

Response: To address this, we have included Figure R2-6, comparing AUC values for PRS models derived from TPMI alone and TPMI+UKB with PRS-CSx. As shown in Figure R2-6, PRS models trained solely in TPMI perform similarly or slightly better for prostate cancer, viral hepatitis B, and type 2 diabetes, while TPMI+UKB-derived models show comparable or improved performance for alcoholism, glaucoma, and migraine. These results highlight population-specific differences in PRS performance and emphasize the importance of ancestry-matched training data. We have incorporated the cross-population PRS into our results, and we suggest this would be a further direction for the PRS development and application.

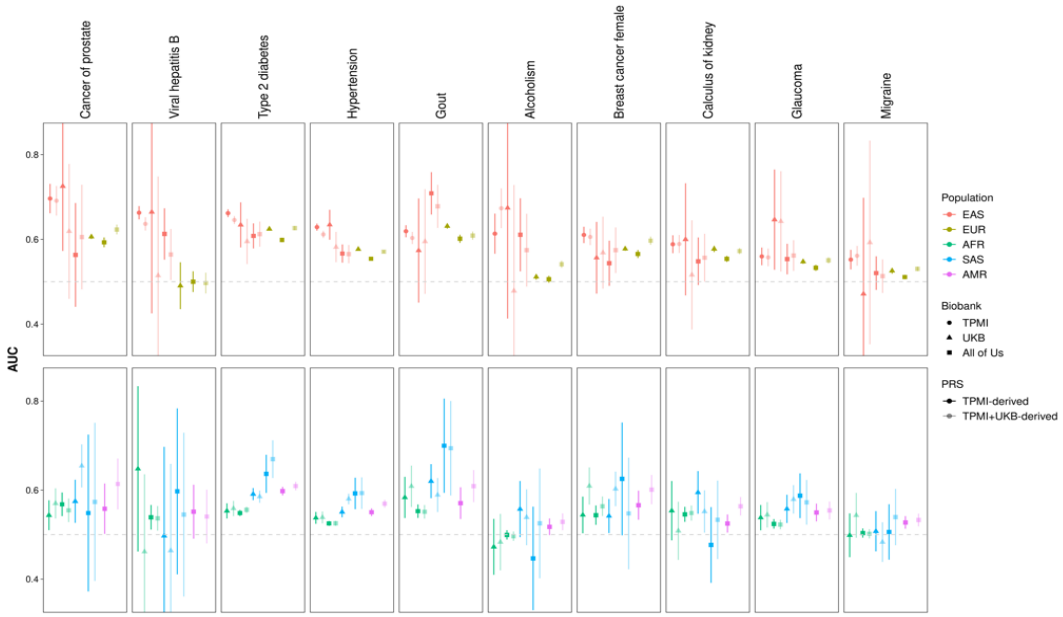


Figure R2-6. (Extended Data Fig. 6 in revised manuscript) External validation of TPMI-derived PRS model and TPMI-UKB cross-population PRS models across populations. PRS validation (AUC ± 95% CI) in East Asian (red), European (olive green), African (green), South Asian (blue), and All of Us (purple) populations from TPMI (circle), UKB (triangle), and All of Us (square) cohorts.

Revised sentences (in results, LINE 612-619)

Additionally, we assessed the performance of TPMI-derived PRS and TPMI-included cross-population PRS across various ancestry groups, including European, African, Admixed American, and South Asian populations from the UK Biobank and All of Us cohorts (Extended Data Fig. 6). The performance varied by diseases, but consistent results were observed for female breast cancer and glaucoma across populations, and TPMI-included cross-population PRS slightly but not significantly improved in the populations other than EAS and EUR.

The Discussion section has been expanded to address PRS transferability across populations.

Revised sentences (in discussion, LINE 707-714)

When comparing with European-derived PRS models, we also observed better performance across several traits in EAS, particularly for cardiometabolic and autoimmune

diseases. Similarly, the TPMI-included cross-population model slightly improves performance in other ancestral populations. These results highlight the need for population-specific prediction models and emphasize the importance of genetic data from diverse populations to advance cross-population models.

Minor

15. 1) It would be helpful to add the correlation line to the plot in Figure 1A to make it consistent with the main text (r=0.64, line 426).

    2) The y-axis for Figure 1B is off (more space between 100-200 than 200-300). This data may be better presented in a table with sample sizes and then mean/median/range or case/control counts to get a sense of scale. This is especially important when considering all the splits of analyses for the PRS development/validation.

    Response: As suggested, we updated Figure 1 as shown below and the additional information were included in revised Table S1 (phecodes) and S2 (quantitative traits).

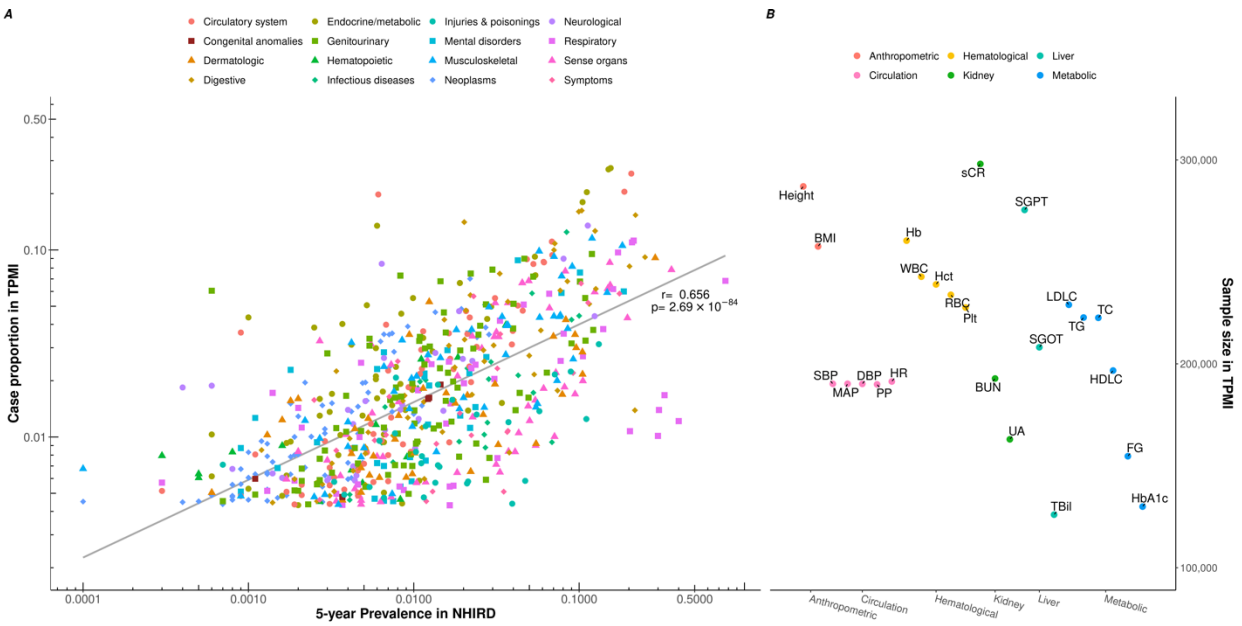


Figure R2-7. (Fig 1 in revised manuscript) Scatter plots of the case proportion and case number in TPMI dataset. (A) The case proportion in TPMI is compared to the 5-year prevalence in the National Health Insurance Research Database (NHIRD) for dichotomized phenotypes (phecodes). Each dot represents a specific phecode, with the x-axis showing the prevalence in NHIRD and the y-axis showing the case proportion in TPMI. (B) Scatter plot shows the sample sizes for quantitative traits in the TPMI cohort. Each point represents a trait, with the x-axis indicating the different category of quantitative traits and the y-axis representing the corresponding sample sizes.

16. The manuscript reads several time as if the investigators are surprised to have a signal for HBV. This is not surprising given the prevalence of this infection and the sequelae associated with chronic infection. If anything, this is an important contribution to the GWAS of infectious disease and speaks to the strength of TPMI re:different epidemiology for specific traits. This could be expanded on a bit more to add a bit of punch to the discussion/conclusions.

Response: We agree with the reviewer's opinion. As suggested, we emphasize the importance of having biobanks or human genetic resources from diverse populations in conditions such as HBV, where the incidence is much higher than that in European populations. The varying environmental exposures and cultural differences among populations may reveal distinct genetic interaction mechanisms between the human genome and external stimuli, which would complete our understanding of human genetics. We have now extended our discussion to give this topic more attention.

Revised sentences (in discussion, LINE 674-681)

Our unexpected success of GWAS and PRS for hepatitis B not only demonstrate the power of the large sample size of TPMI, but also reveal the necessity of population-specific genetic study for population-unique diseases. The benefits extend beyond differences in ancestry to include environmental factors such as pathogen exposure, food intake, and lifestyle influences. Exploring how the human genome interacts with these diverse external and environmental factors can significantly enhance our understanding of how genetic variants contribute to disease susceptibility or severity.

17. Why were the specific cutoffs for PRS used in lines 522-524? For example, why look for models explaining 3% of the trait variance? Or 0.55 for the AUC?

(Polygenic Risk Score (PRS) Development) Out of the 289 PRS models, AUC values surpassed 0.55 for 106 dichotomized phecodes, while all 24 models for quantitative traits accounted for more than 3% of the phenotypic variance.

Response: Our original intention was to report the number of PRS models with significantly good performance (p-value < 0.05). However, p-value alone is not an ideal index for evaluating performance. Therefore, we examined the relationship between performance indices (AUC and R^2) and p-value. We found that AUC > 0.55 is the thresholds that assure models with a significant p-value for phecodes' PRS. And all quantitative traits show a R^2 with significant p-value That's the rationale behind these arbitrary thresholds, and we have now included the p-value threshold in the revised statement.

Revised sentences (in results, LINE 549-553)

Out of the 265 PRS models for phecodes, AUC values exceeded 0.55 for 105 dichotomized phecodes with a significant p-value (p-value < 0.05). Additionally, the explained variance of models for 24 quantitative traits ranged from 0.028 (aspartate aminotransferase, AST) to 0.227 (height). (Table S9 and Extended Data Fig. 5).

18. Please remove AUC from the plots showing r2 and heritability for PGS. They are not comparable and the plots are misleading, especially since the AUC is often way above the h2 (because they are not the same). AUC can be a separate plot.

Response: As suggested, we have updated figure 5 and Extended Data Fig 5 as following:

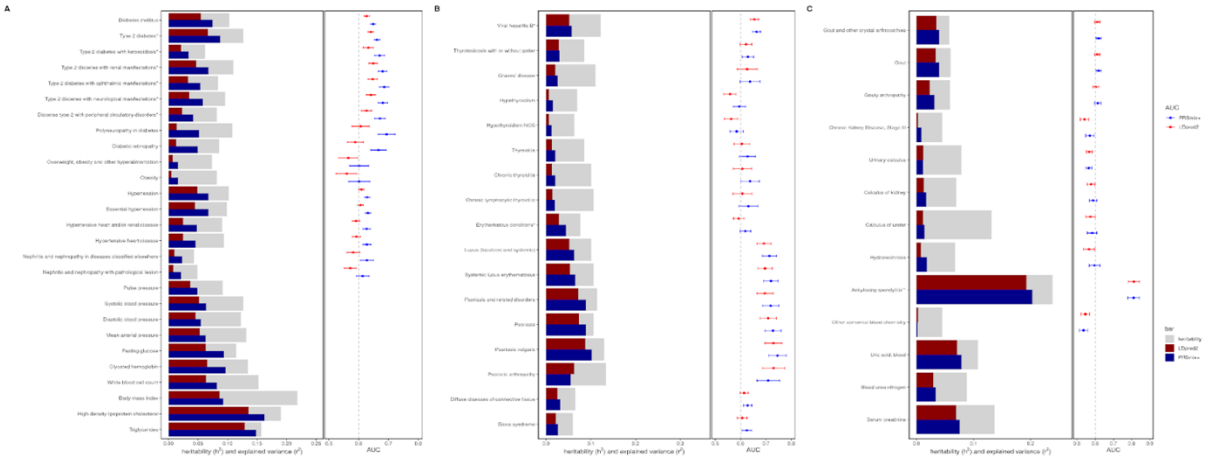


Figure R2-8 (Fig. 5 in revised manuscript) PRS performance for the three identified trait clusters. Bar plot shows SNP-heritability and PRS explained variance (r^2) for (A) cardiometabolic trait cluster, (B) autoimmune trait cluster, and (C) kidney-related trait cluster. Gray bars indicate SNP-heritability, and the colored bar chart presents the r^2 values, indicating the proportion of variance explained by the PRS from single-trait PRS (LDpred2, red bar) or multi-trait PRS (PRSmix+, blue bar), while the dot and whisker plot showcases predictive accuracy using AUC. The area under the receiver operating characteristic curve (AUC) presented with 95% confidence interval for dichotomized

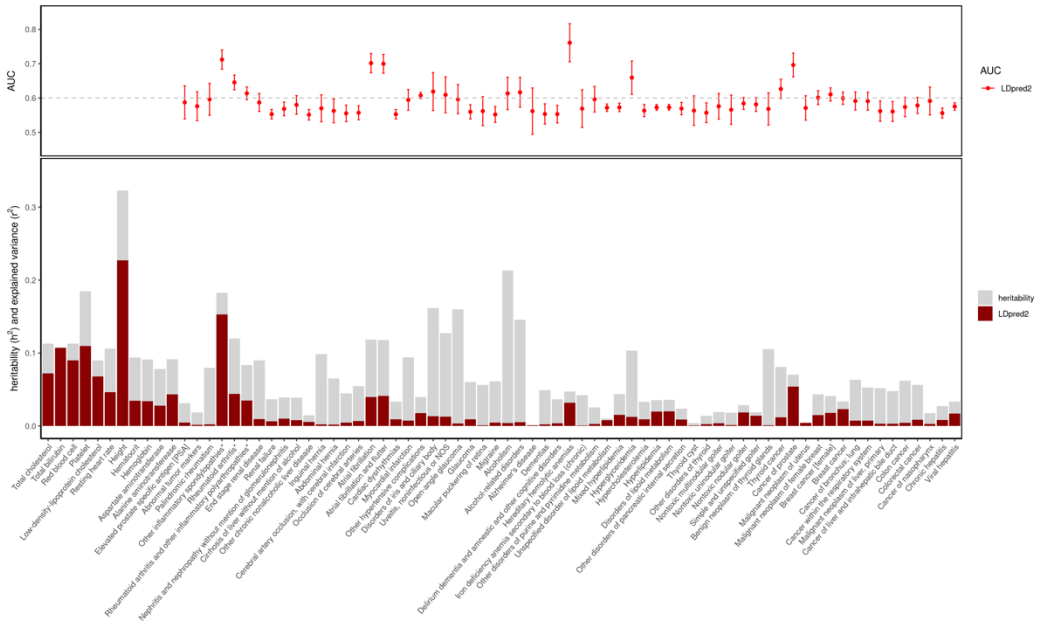


Figure R2-9. (Extended Data Fig. 5 in revised manuscript) The bar chart and dot plot for PRS performance. Bar and dot plot showing PRS explained variance (r^2) and SNP-heritability for dichotomous traits. Gray bars represent SNP-heritability, and dots show AUC. An asterisk (*) indicates estimates considering the MHC region

19. Is the use of "epidemic" on line 616 a typo? Did you perhaps mean "endemic"? (I would also expand in the discussion as to why geographically and ancestrally diverse biobanks are needed, as the epidemiology of HBV is very different in Taiwan versus other places.)

    (**Discussion)** Our understanding of how the genetic factors influencing Hepatitis B, an epidemic infectious disease in Taiwan with an estimated hepatitis B virus carrier rate of 15-20%, also benefited from the large dataset.

    Response: “Epidemic” on line 616 was a typo. It was an “endemic” condition previously, but, with a very successful vaccination program, it is better characterized as a prevalent infectious disease.

    Revised sentences (in discussion, LINE 666-681)

    Our understanding of how the genetic factors influencing Hepatitis B, an endemic infectious disease in Taiwan with an estimated hepatitis B virus carrier rate of 9.78% among the unvaccinated cohort (born before 1984)[28], also benefited from the large dataset. With 23,618 cases, a significant increase from prior studies of only a few thousand cases[37-39], we identified novel loci and demonstrated that host genome may determine the severity and symptoms of this infectious disease. This is similar to that previously reported in COVID-19 and pneumonia, where genetic factors have been shown to influence disease outcomes[40-43]. Our unexpected success of GWAS and PRS for hepatitis B not only demonstrate the power of the large sample size of TPMI, but also reveal the necessity of population-specific genetic study for population-unique diseases. The benefits extend beyond differences in ancestry to include environmental factors such as pathogen exposure, food intake, and lifestyle influences. Exploring how the human genome interacts with these diverse external and environmental factors can significantly enhance our understanding of how genetic variants contribute to disease susceptibility or severity.

20. Another source of diverse eQTL data could be MAGE (https://github.com/mccoy-lab/MAGE), just in case you hadn't seen that yet.

    Response: We now incorporate two additional eQTL resources with our colocalization analysis, the multi-ancestry lymphoblastoid cell lines eQTL, MAGE, and Japanese whole blood eQTL, JCTF. These two resources revealed 309 additional gene-trait pairs (coloc H4>0.9) comparing to GTEx whold blood eQTL, including 153 pairs from MAGE and 191 pairs from JCTF. We have revised our result and discussion to emphasize the importance of diverse populations for other type of genetic dataset based on these findings.

    Revised sentences (in results, LINE 502-510)

    We also conducted a colocalization analysis to elucidate the potential molecular function of identified GWAS signals with three QTL datasets, including GTEx[32], MAGE[33], and JCTF[34]. (Fig. 3 and Table S8). Our results identified 391 unique genes that significantly mediate the outcome through their expression (posterior probability > 0.9), including *GBAP1* which colocalized with five different traits (uric acid, serum creatinine, hematocrit, hypertension, and gout). Among the colocalized genes, 75 of them can be identified only in the multi-ancestry lymphoblastoid cell lines eQTL, MAGE (20 genes), and/or Japanese whole blood eQTL, JCTF (59 genes).

Revised sentences (in discussion, LIN 750-758)

Second, we attempted to use eQTLs to elucidate the molecular mechanism of diseases, but the underrepresentation of EAS in current eQTL datasets, such as GTEx, poses challenges[32]. Gene expression regulation varies across ancestries[44,45], and differences in LD structures further complicate colocalization analyses. Comparing to GTEx whole blood eQTL, the multi-ancestry lymphoblastoid eQTL and Japanese whole blood eQTL revealed 309 additional gene-trait pairs. Therefore, ancestral diversity is an urgent need not only in genomic data but also in transcriptomic, proteomic, metabolomic, and epigenomic datasets.

Referee #3 (Remarks to the Author):

Chen et al. present the Taiwan Precision Medicine Initiative (TPMI), which is a truly remarkable dataset. I enjoyed reading the manuscript, and seeing how far the genetics field has come. TPMI includes over half a million Taiwanese residents, making it possibly the largest east asian genetic dataset with electronic medical record data. I think this is a very impressive study, which addresses a critical gap in precision medicine for non-European populations, and will likely have a very significant impact on the human genetics field. The publicly available online resource and GWAS summary statistics will likely be widely used in future genetic studies. There are many interesting analyses in the paper, and generally I found them to be of high quality, e.g. GWAS and PGS analyses were impressive. It is also a very well written paper. However, I suggest reconsidering the emphasis on predicting disease burden in Taiwan, as I believe selection biases may significantly impact these numbers, possibly rendering them meaningless. Besides this comment, most of my other comments are relatively minor, and should be straightforward to address in a revised manuscript. The comments below are aimed at improving the manuscript.

Response: We thank the reviewer for the positive feedback and appreciate the suggestions that help us improve our manuscript. We have embraced the suggestions and revised the paper accordingly.

Regarding the concern about potential selection biases in the disease burden analyses, we have revised the manuscript to address this limitation. We now frame these analyses as exploratory, highlighting their purpose in demonstrating the integration of genomic data with healthcare utilization trends rather than providing definitive estimates. The text explicitly acknowledges the influence of TPMI's volunteer-based recruitment and demographic characteristics, such as overrepresentation of middle-aged participants and urban populations.

Revised sentences (in discussion, LINE 723-736)

TPMI's case proportion also shows a moderate but significant correlation with the prevalence from National Health Insurance dataset, and this may imply the ascertainment bias of TPMI's hospital-based design. Compared to the general population in Taiwan (NHIRD), TPMI participants are overrepresented in the middle-aged group (year of birth 1940-1970, 54.3% vs. 38.3%), include slightly more females (55.1% vs. 50.6%), and have a higher proportion of participants from northern Taiwan (59.5% vs. 47.3%). These demographic differences, along with the volunteer-based recruitment process, likely contribute to the lower case proportions observed in TPMI (Figure 1a). Importantly, a significant portion of disease records in the NHIRD originate from local clinics and primary care settings, which are not covered in TPMI. Notably, the EMR of the participants are incomplete, as some participants receive care from multiple health providers, but the TPMI only has access to EMRs from their enrollment hospitals.

1. I would appreciate at least some comments on selection biases, and how they might bias the analyses, and perhaps consider employing approaches to address these. These could include inverse probability weighting (see e.g. Schoeler et al. Nat Hum Behav 2023.

Response: Selection biases are important concerns and appreciate the suggested methods to address them. We have revised the manuscript to explicitly acknowledge and discuss these potential biases. Specifically, we note that TPMI's volunteer-based recruitment and demographic characteristics, including the overrepresentation of middle-aged and urban participants, may affect the generalizability of the findings.

Regarding the suggestion to employ approaches like inverse probability weighting to mitigate these biases, we are unable to acquire at the present time the external reference data needed to calculate accurate probabilities for weighting across all traits. Nonetheless, we have included a detailed discussion of potential biases and their implications in the revised text and clarified the exploratory nature of the disease burden analysis. We will actively strive to collect external reference data in the future studies to formally address this important issue.

Furthermore, we have added the paper by Schoeler et al.[4] to our references in the revised manuscript to highlight methodological considerations for addressing selection biases in future studies.

Revised manuscript (in Discussion, LINE 736-746)

We acknowledge that ascertainment biases may influence disease prevalence and heritability estimates. Thus, we accounted for case-control ascertainment in heritability estimates and PRS evaluations by applying liability-scale transformations using population prevalence data from NHIRD and utilized independent validation datasets for PRS validation (TWB, UKB, All of Us) to mitigate the impact of ascertainment biases, while acknowledging that residual biases may persist. Methods like inverse probability weighting could mitigate such biases[4], but these require detailed external reference data that are currently unavailable. This limitation also emphasizes the need for future adjustments to enhance generalizability.

1.a). This study evaluated the "genetic impact on overall disease burden". I believe this estimate could be very biased due to selection biases, including participation bias, as well as time-related biases (e.g. right and left censoring).

Response: We agree that selection and time-related biases may impact our estimate of genetic contributions to disease burden. We have revised the manuscript to acknowledge these biases explicitly and to clarify the limitations of our analysis, including the influence of right and left censoring. These updates emphasize the exploratory nature of our findings and the need for refined methods in future studies.

Revised sentences (in discussion, LINE 716-721)

Although the estimates of genetic contributions to health measure may be influenced by selection biases, including participation bias and time-related biases such as right and left censoring, our result suggest integrating genetic risk-based health management strategies with traditional risk factors, such as age, sex, smoking status, and BMI, may enhance prediction models and refine personalized risk stratification.

1.b) Similarly, heritabilities and PGS accuracies will likely be impacted by selection biases. Interestingly, you do adjust for case-control ascertainment. Perhaps you could comment on this?

Response: As noted in our response to Reviewer #2, we accounted for case-control ascertainment in our heritability estimates by applying liability-scale transformations based on population prevalence data from NHIRD. This adjustment helps mitigate the bias introduced by oversampling cases in the TPMI cohort.

Similarly, for PGS evaluations, we utilized independent validation datasets to reduce the potential confounding effects of selection biases within TPMI. However, we acknowledge that some residual biases may remain due to unmeasured factors. We have clarified these points in the revised manuscript and highlighted the steps taken to address selection biases where possible.

Revised sentences (in discussion, LINE 738-742)

We accounted for case-control ascertainment in heritability estimates by applying liability-scale transformations using population prevalence data from NHIRD and utilized independent validation datasets for PRS evaluations (TWB, UKB, All of US) to mitigate the impact of selection biases, while acknowledging that residual biases may persist.

2. I was impressed by the large number of PGS methods that you used. However, I would have liked to see a clearer comparison between the different PGS methods. You claim that LDpred2 outperformed other methods, but it's hard to see from the supplementary tables.

Response: While PRS method comparison was not the main thrust of our study, we agree that presenting a more transparent comparison will improve clarity. In the revised manuscript, we have added a summary figure to the main text and reorganized the supplementary tables to explicitly highlight the comparative performance of the PRS methods across key traits. (Figure R3-1, the newly added Extended Data Fig 4. in the revised manuscript) This figure illustrates the superior performance of LDpred2 for most traits, as reflected in higher AUC values, as compared to other methods. These updates improve transparency and clarify the basis for selecting LDpred2 as the

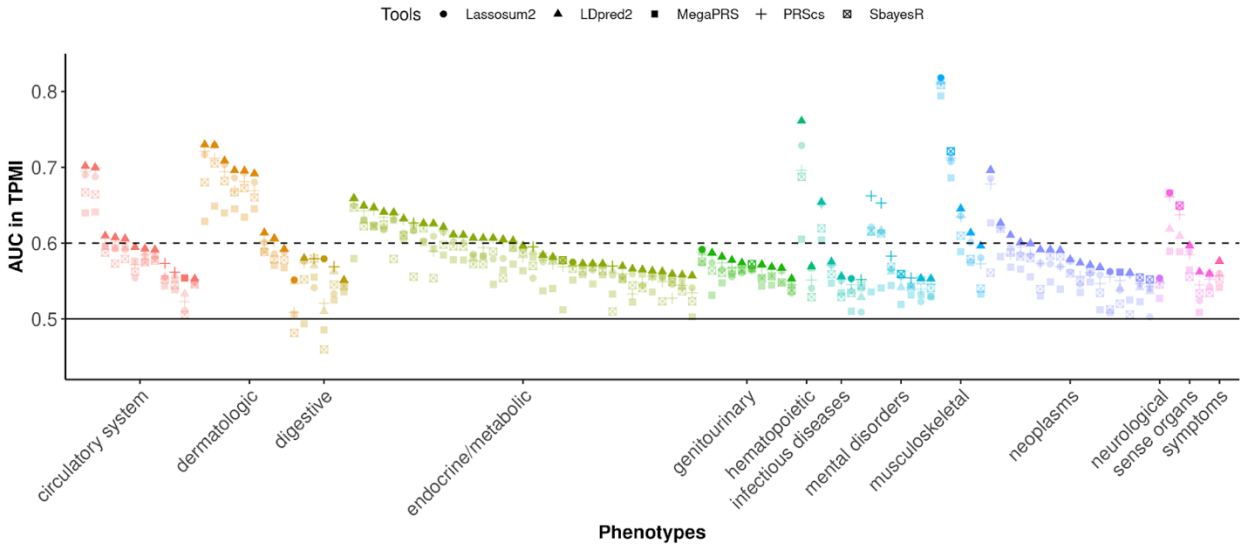


Figure R3-1. (Extended Data Fig 4. in revised manuscript) The scatter plot of PRS performance of PRS tools. This figure illustrates the superior performance of LDpred2 for most traits, as reflected in higher AUC values, as compared to other methods

primary tool for further analyses. We hope this addition provides a clearer comparison and strengthens the manuscript.

3. What does “Researchers requesting access to the individual genotyping and EMR data can do so on a collaborative basis.” mean? This statement seems rather ambiguous to me, and I think it could be improved.

Response: The data availability statement was meant to show our eagerness to collaborate with interested researcher while adhering to strict Taiwanese HIPAA laws that govern use of clinical and genetic data. New laws and regulations are being drafted by the Taiwan government to allow large datasets such as that of the TPMI to distribute de-linked individual data to researchers anywhere around the world. Before the passage of this law and reconfiguration of the TPMI dataset (a process that will take about 18 months), the only way for researchers outside the TPMI Consortium to access the data is through collaboration. We have worked with the Nature Editorial Office to revise the data availability statement to comply with Nature’s data sharing standards and Taiwan government’s regulation.

4. In line 645 you say that “implementing genetic risk-based health management strategies may decrease the disease burden”. Although I don’t necessarily disagree, and I don’t know what the future holds, I would be surprised if the genetic predictors radically improved prediction of disease burden compared to standard risk factors (e.g. age, sex, smoking status, BMI, etc.)

Response: We agree that traditional risk factors such as age, sex, smoking status, and BMI currently play a significant role in predicting disease burden. Our intention was not to imply that genetic predictors alone would radically outperform these factors but rather to highlight the complementary value of integrating genetic risk with standard risk factors to enhance prediction models. However, for a small subset of patients with high genetic risk (e.g., those in the top 5% of PRS score for a disease), early screening for cancer or early intervention for metabolic diseases will likely play a role in decreasing overall disease burden in the population while improving health outcomes in the high-risk individuals.

To clarify this point, we have revised the manuscript to temper the language, emphasizing that genetic predictors are most impactful when combined with established clinical and lifestyle risk factors, potentially refining risk stratification and personalized health management strategies.

Revise sentences (in discussion, LINE 716-721)

Although the estimates of genetic contributions to health measure may be influenced by selection biases, including participation bias and time-related biases such as right and left censoring, our result suggest integrating genetic risk-based health management strategies with traditional risk factors, such as age, sex, smoking status, and BMI, may enhance prediction models and refine personalized risk stratification.

5. In line 764-6 you note that you used “PLINK2 for GWAS among the unrelated subset (n = 248,754)[51], and these statistics were used for heritability and genetic correlation estimation.” I hope you also used these for the PGS, as LDpred2 and most other methods assume the GWAS summary statistics are obtained using a logistic regression or linear regression, and not from mixed models (e.g. BOLT-LMM and SAIGE). If not, then please update this, as I am confident that your PGS results will improve.

Response:

To ensure methodological consistency, we tested PRS models using summary statistics from both logistic regression models (PLINK with the unrelated set, n=248,757) and mixed-effects models (SAIGE with full GWAS set, n=363,477). As shown in Figure R3-2, the PRS models built using GWAS summary statistics from mixed-effects models (SAIGE) performed slightly better than those built using statistics from logistic regression models (PLINK2 on unrelated individuals). However, this difference was not statistically significant across the tested traits. Since we used independent and unrelated individual-level data for PRS model fine-tuning and parameter selection, we were able to leverage the larger sample size of mixed-effects models (which improves GWAS power) while minimizing potential biases in PRS estimation. Given the above results, we retained SAIGE-based GWAS summary statistics for PRS construction.

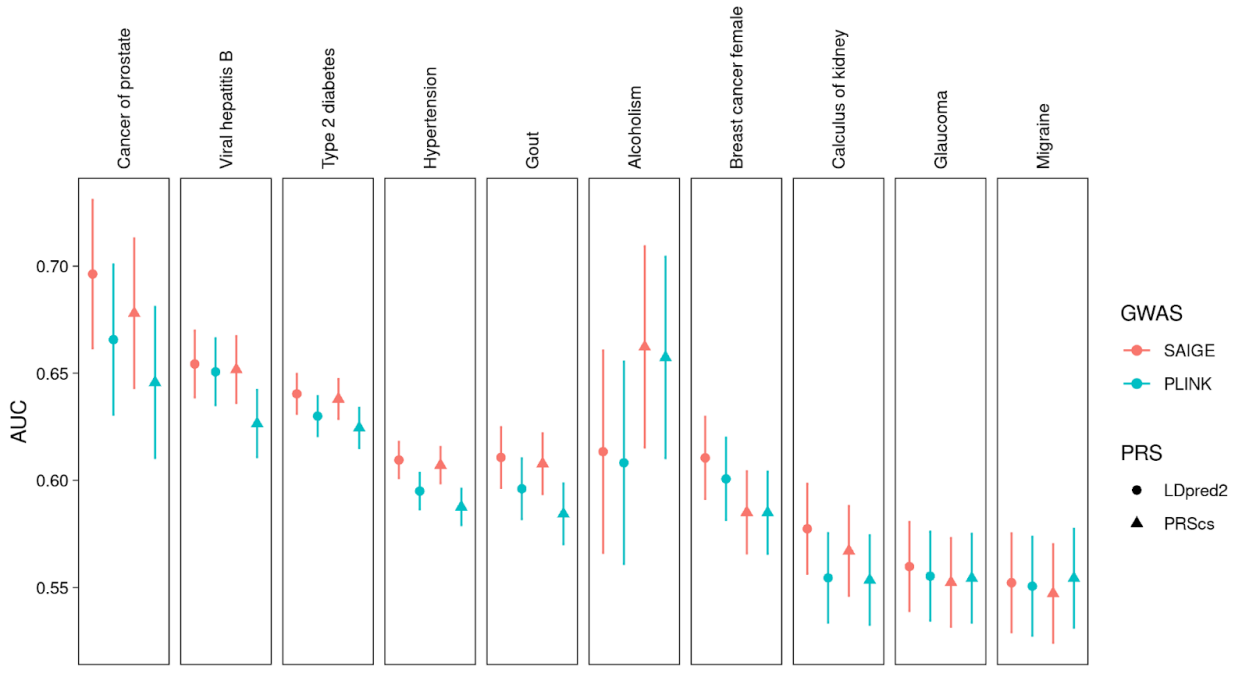


Figure R3-2. The PRS performance comparison for the models using GWAS summary statistics, logistic regression from PLINK (blue) and mixed -effect model (red).

6. You use SuSiE to finemap the variants, using LD derived from the imputation reference panel. I worry that this imputation panel might not capture the LD in the GWAS sample (with imputed SNPs). I believe you could improve the fine-mapping by switching to the LD between the (imputed) SNPs in the GWAS sample, although this will likely not change much in practice.

Response: Using LD derived directly from the GWAS sample with imputed SNPs could, in theory, provide a more precise representation of the LD structure specific to our study cohort. However, our imputation reference panel is highly concordant with the genetic architecture of our study population, minimizing potential discrepancies. To assess the impact of using LD from different sources, we compared fine-mapping results from nine randomly selected regions using both approaches (Figure R3-3). The results showed nearly identical causal variant posterior

probabilities and credible set compositions, indicating that the use of the imputation reference panel does not meaningfully affect fine-mapping accuracy in our dataset.

Additionally, using LD from the full GWAS sample significantly increased computational time (15-60 minutes vs. 10 seconds per fine-mapped region), making it less practical for large-scale analyses. Given the minimal difference in results, we retained the imputation reference panel LD approach to balance efficiency and accuracy. We have noted this as a potential refinement in the revised manuscript, emphasizing that while the current approach is robust, future analyses could incorporate LD calculated directly from the GWAS sample to enhance precision. We appreciate your suggestion and its potential to improve fine-mapping accuracy.

Revised sentences (in methods, LINE 900-904)

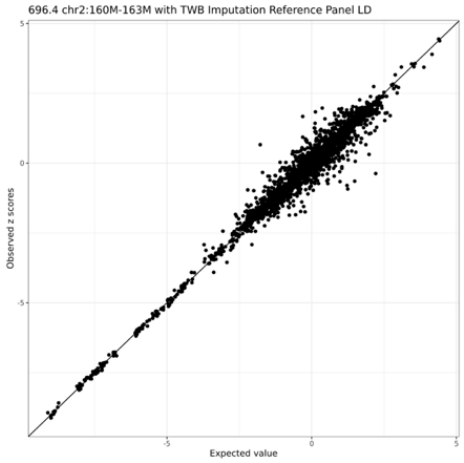
SuSiE was conducted for this summary statistics-based fine-mapping with LD derived from our imputation reference panel[46], which reflects our study population's genetic architecture. While using LD from the GWAS sample could improve accuracy, we used the imputation panel due to the computation efficiency.

Figure R3-3. The diagnostic plots of SuSiE fine-mapping with LD from imputation reference (A, C, E, G, I, K, M, O, Q) and from GWAS samples (B, D, F, H, J, L, N, P, R)

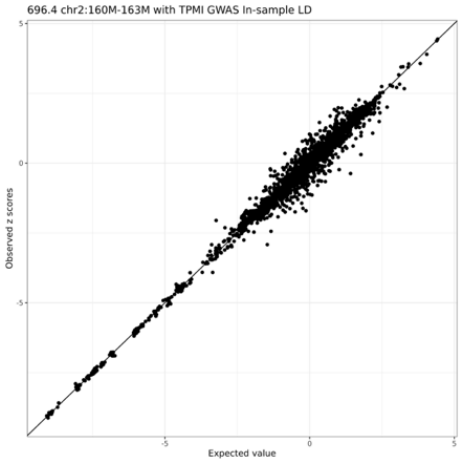
LD from imputation reference

LD from GWAS samples

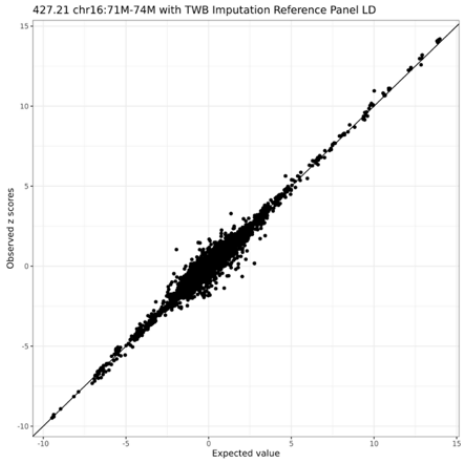
A)



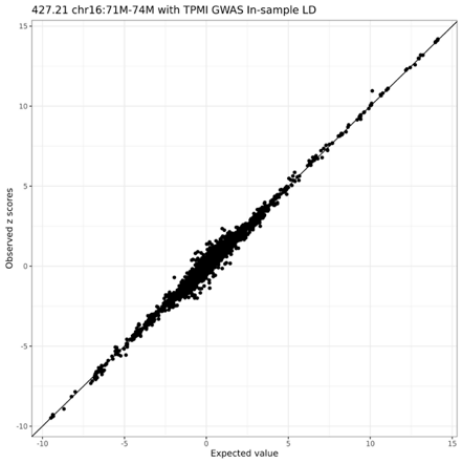
B)



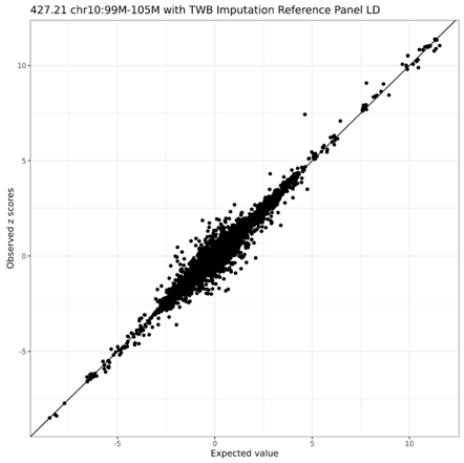
C)


427.21 chr16:71M-74M with TWB Imputation Reference Panel LD
Observed z scores
Expected value

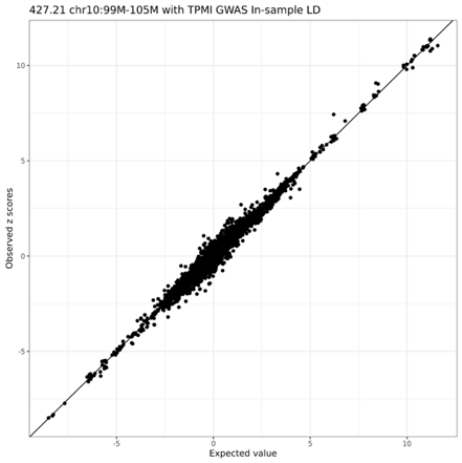
D)


427.21 chr16:71M-74M with TPMI GWAS In-sample LD
Observed z scores
Expected value

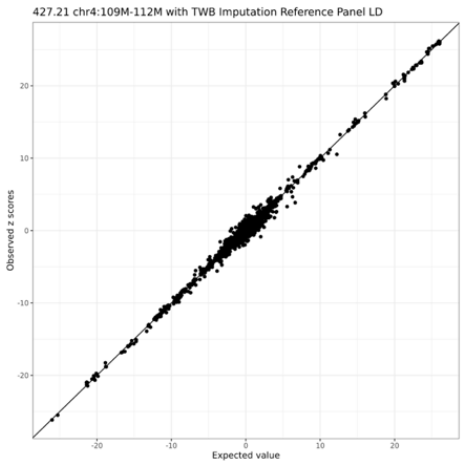
E)


427.21 chr10:99M-105M with TWB Imputation Reference Panel LD
Observed z scores
Expected value

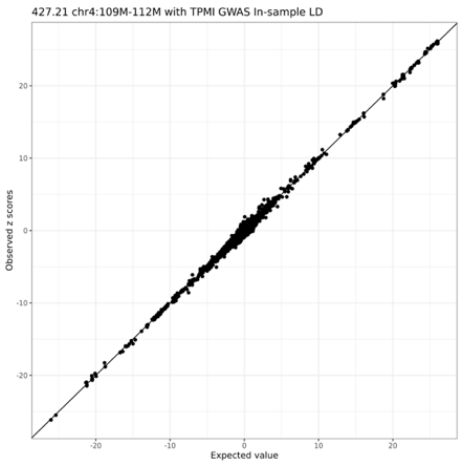
F)


427.21 chr10:99M-105M with TPMI GWAS In-sample LD
Observed z scores
Expected value

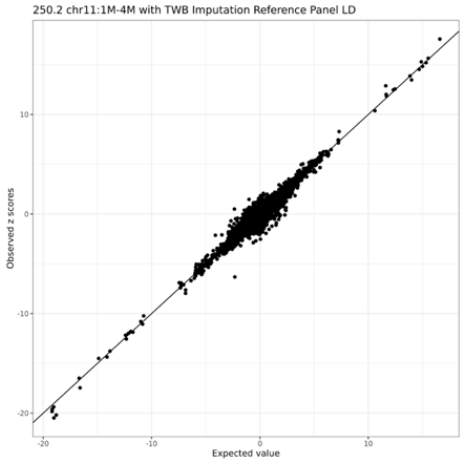
G)


427.21 chr4:109M-112M with TWB Imputation Reference Panel LD
Observed z scores
Expected value

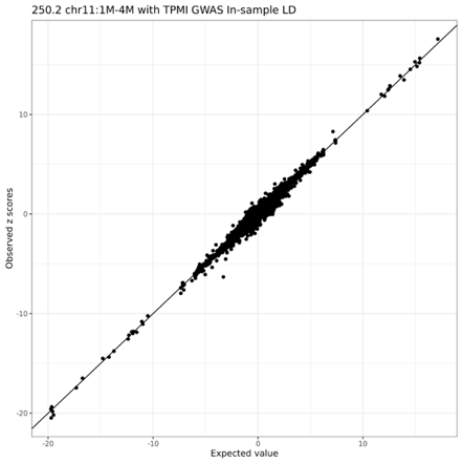
H)


427.21 chr4:109M-112M with TPMI GWAS In-sample LD
Observed z scores
Expected value

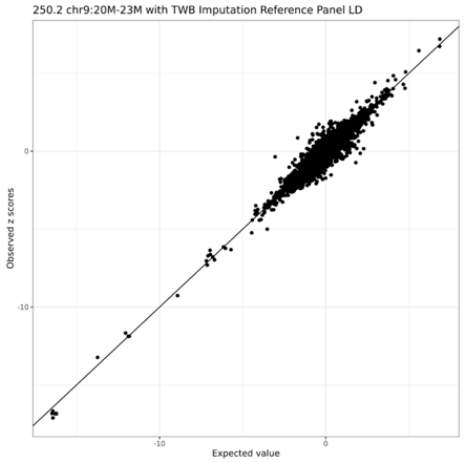
I)


250.2 chr11:1M-4M with TWB Imputation Reference Panel LD
Observed z scores
Expected value

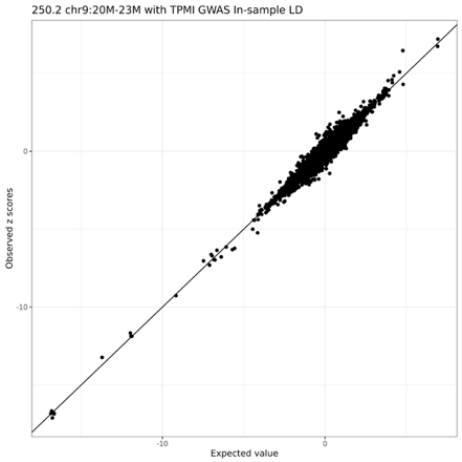
J)


250.2 chr11:1M-4M with TPMI GWAS In-sample LD
Observed z scores
Expected value

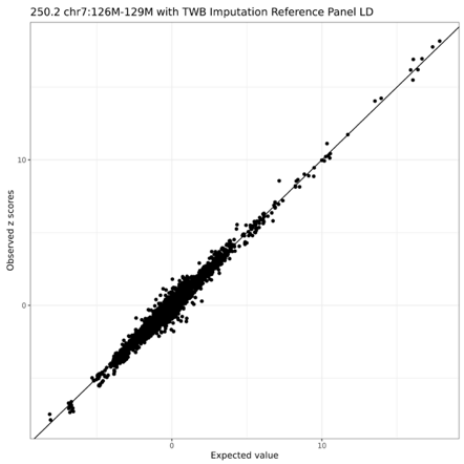
K)


250.2 chr9:20M-23M with TWB Imputation Reference Panel LD
Observed z scores
Expected value

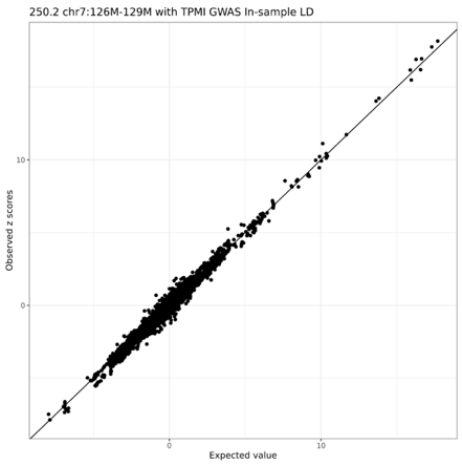
L)


250.2 chr9:20M-23M with TPMI GWAS In-sample LD
Observed z scores
Expected value

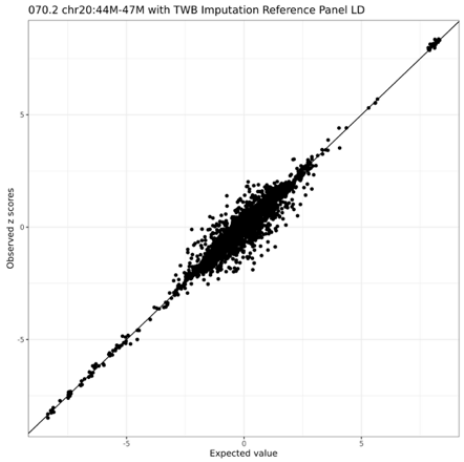
M)


250.2 chr7:126M-129M with TWB Imputation Reference Panel LD
Observed z scores
Expected value

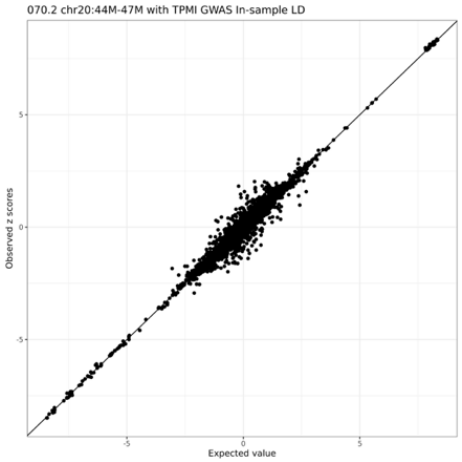
N)


250.2 chr7:126M-129M with TPMI GWAS In-sample LD
Observed z scores
Expected value

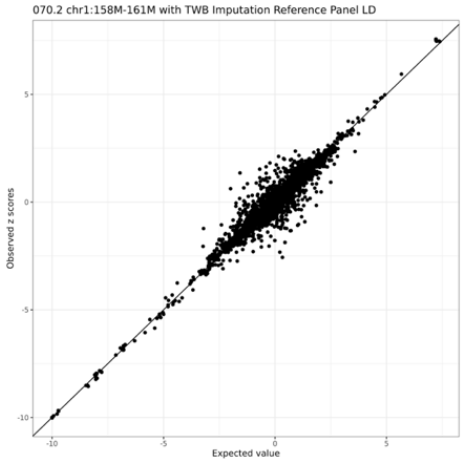
O)



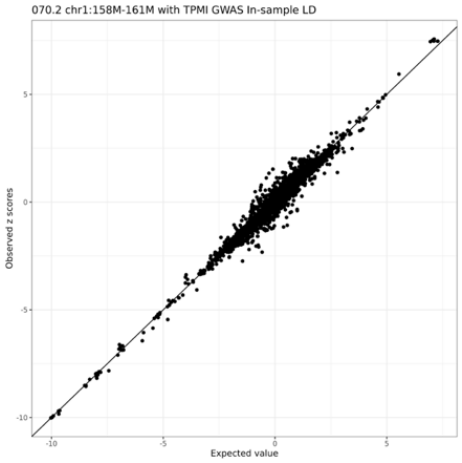
P)



Q)



R)



7. The description of the novel variant identification procedure is unclear to me. You write (1) “We classified a gene or region as novel if the fine mapped independent signal was not located within 1 Mb of any reported genome-wide significant association (p-value < 5x10-8) for the corresponding phenotype55.” and (2) “Additionally, a variant was considered a novel hit if the highest linkage disequilibrium (LD) r² was less than 0.1 with any reported significant association.” Do you require both conditions, (1) and (2), to be true or is (2) enough to say the variant is novel?

Response: To clarify this important point, we tried to describe two kinds of novel findings (variants in regions not previously associated with the condition and those in known regions of association but not in LD with the associated loci). Given the confusion, we have now switched to call them “novel variants” and “new hits”. The “novel variant” meets the first requirement: “We classified a gene or region as novel if the fine-mapped independent signal was not located within 1 Mb of any reported genome-wide significant association (p-value < $5x10^{-8}$) for the corresponding phenotype”” Also, the “new hit” meets the second requirement: “a variant was considered a new hit if the highest linkage disequilibrium (LD) r² was less than 0.1 with any reported significant association.”

We have revised the manuscript to explicitly state this criterion to avoid ambiguity and ensure clarity. Thank you for pointing out the need for this clarification.

Revised sentences (in results, LINE 441-446)

Notably, 95 novel associations, defined as having no previously reported results within 1Mb in the NHGRI-EBI GWAS Catalog[47] of human GWAS, were identified across 50 phecodes and 7 quantitative traits. In addition, we identified 217 new hits from previously reported regions, defined as having r^2<0.1 with any variant observed in the NHGRI-EBI GWAS Catalog within 1 MB for the same phenotype (Table S4 and S5).

Revised sentences (in methods, LINE 909-914)

We classified a variant as novel if the fine mapped independent signal was not located within 1 Mb of any reported genome-wide significant association (p-value < $5x10^{-8}$) for the corresponding phenotype. Additionally, a variant was considered a new hit if the highest linkage disequilibrium (LD) r^2 was less than 0.1 with any reported significant association within 1Mb.

8. You note that you use LDSC to estimate the heritability. However, I wonder how this estimate compares with the LDpred2-auto estimate, which assumes a point-normal prior, and could be much more accurate for outcomes with relatively simple genetic architectures, e.g. lipid measurements. Alternatively, you could use SBayesS to estimate the heritability.

Response: While we primarily employed LDSC for heritability estimates, we recognize that methods such as LDpred2-auto, which assumes a point-normal prior, or SBayesS, may offer complementary insights, especially for traits with simpler genetic architectures, such as lipid measurements. To evaluate potential differences, we compared SNP-heritability estimates from LDSC, LDpred2-auto, and SBayesS across 24 quantitative traits (Figure R3-4). The results indicate that heritability estimates from LDpred2-auto are similar to those from LDSC, whereas estimates from SBayesS tend to be lower for most quantitative traits. Given that LDSC is widely used for heritability estimation, we decided to report the LDSC estimates to facilitate comparability with other studies. Furthermore, since we also provide the PLINK GWAS summary statistics, they can be readily utilized for heritability estimation with LDpred2-auto and SBayesS.

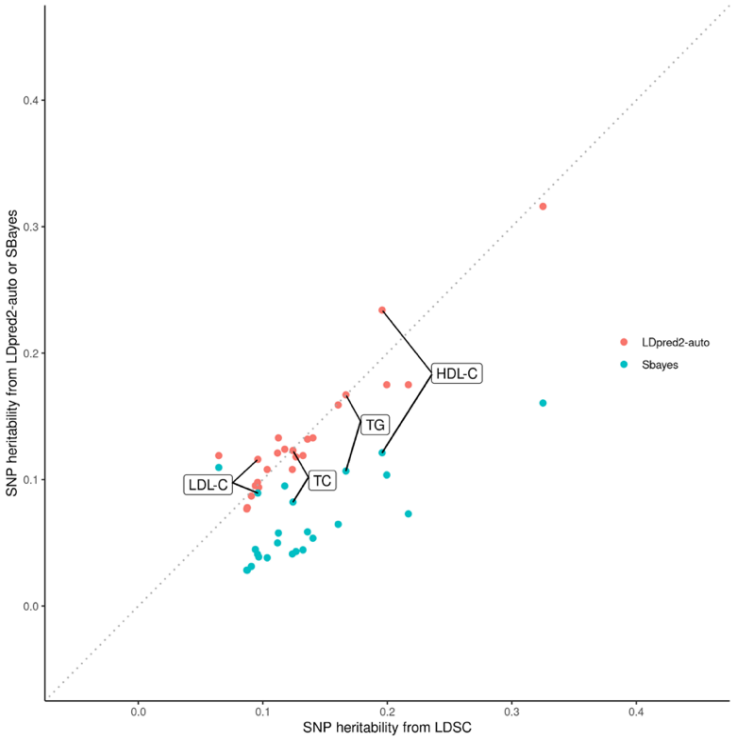


Figure R3-4. The scatter plot of comparison of heritability estimation between LDSC (x-axis) and LDpred2-auto (red) and SBayesS (blue)

9. It was unclear to me how you converted the r2 to liability scale. Was it using the same equation as the heritability, or something like Lee et al. (Gen Epi 2012)?

Response: We performed the liability-scale transformation using the same approach applied for heritability, incorporating the observed r2, the population prevalence of the trait (from NHIRD), and the sample prevalence in the cohort. This approach aligns with the framework outlined in Ojavee et al.[48], which accounts for case-control ascertainment and ensures consistency between heritability and variance explained metrics.

For the traits with higher sample prevalence than population prevalence, the equation from Lee et al. was applied[35]:

$$h^2_{lib} = h^2_{obs} \frac{K(1-K)}{\varphi(\Phi^{-1}(K))^2} \frac{K(1-K)}{P(1-P)}$$

where $\mathrm{h}^2_{\mathrm{obs}}$ is the observed heritability, K is the population prevalence from population, P is the sample prevalence, and $\varphi(\Phi^{-1}(\mathrm{K}))^2$ is the squared probability density function of the standard normal distribution evaluated at the Kth quantile of the inverse cumulative density function of the standard normal distribution.

For the traits with lower sample prevalence than population prevalence, the following equation from Ojavee et al was applied.[48]:

$$h^2_{lib} = h^2_{obs} \frac{(h^2_{lib}\varphi\left(\Phi^{-1}(K)\right)^2 + K - \widetilde{\Phi}(\Phi^{-1}(K), \Phi^{-1}(K), h^2_{lib})^2}{\varphi\left(\Phi^{-1}(K)\right)^2 P(1-P)}$$

We have clarified this process and included the appropriate reference in the revised manuscript.

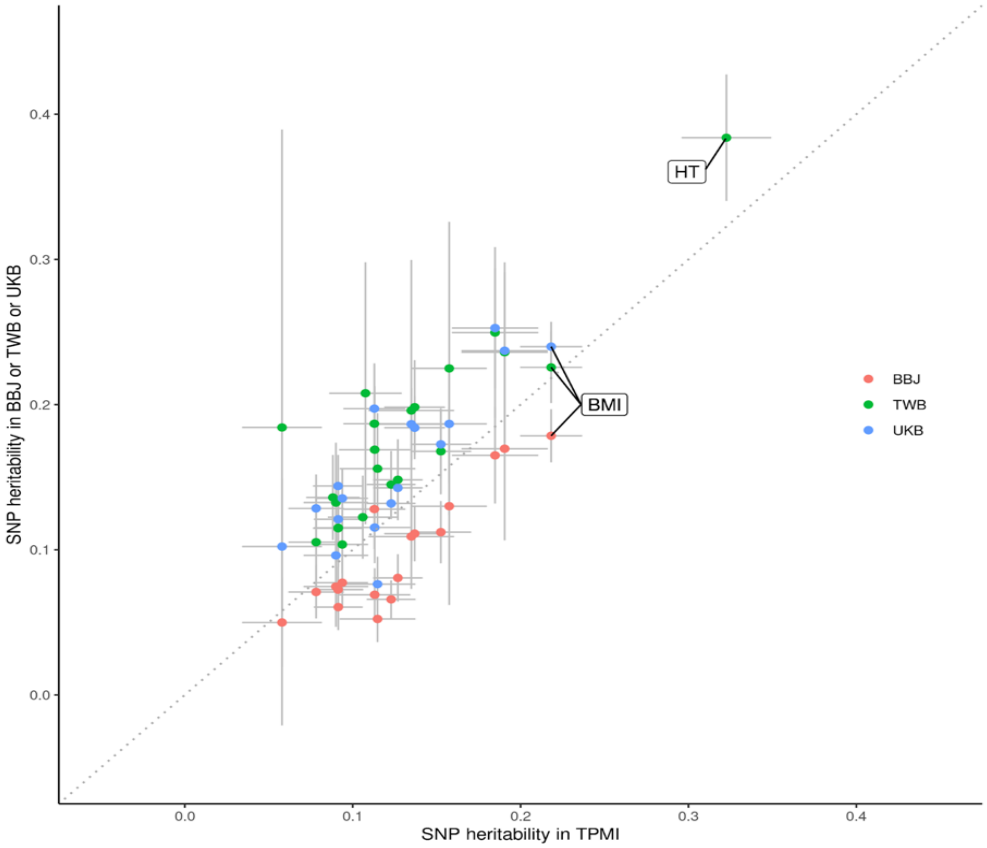
Revised sentences (in methods, LINE 926-932)

For the dichotomized traits, we performed a liability-scaled transformation on the observed heritability using the 5-years population prevalence from the National Health Insurance dataset (NHIRD) of the Health and Welfare Data Science Center[3,48]. For the traits with a higher prevalence in our dataset (TPMI) than the population (NHIRD), we applied the equation derived by Lee et al[35]. For other traits, we used the adapted equation from Ojavee et al[48].

10. I am surprised that height has h2 of only 0.323 and BMI a h2 of 0.218. These seem small compared to the literature. Although this could very well be possible, it makes me wonder whether the GWAS was adjusted for sex. (Perhaps this relates to the following comment?)

Response: As we addressed in our response to Reviewer #2 (comment 6), who asked a similar question regarding height and BMI, the heritability estimates were derived from GWAS adjusted for key covariates, including sex, age, age², and their interactions with sex, to account for potential confounding. We compared our heritability estimates with those from the Taiwan Biobank, Biobank Japan, and UK Biobank[12], which used the same statistical approach (Figure R3-5). The results suggest that our estimates are consistent with previously reported heritability. Although the heritability of body height is slightly lower than the estimate from the Taiwan Biobank, this may be explained by the fact that the EMR-extracted body height includes both self-reported and measured values, which may introduce more phenotypic variation compared to the Taiwan

Biobank, where measurements were taken by a team of trained researchers following the same protocol.



FigureR3-5. The estimates of heritability for 24 quantitative traits in TPMI (x-axis) and other biobanks (y-axis), including Taiwan Biobank (TWB, green), Biobank Japan (red), and UK Biobank (blue).

11. In the validation of PRS, it seems that you don't adjust for anything. Although you've already accounted for sex, age, chip, PCs, etc., in the GWAS, this can be an issue when comparing the r2 to h2, and when predicting into other cohorts. E.g., if you estimate h2 for height after adjusting for sex, then you should also adjust for sex when estimating the PGS r2, otherwise they won't be on the same scale. Also, when using UKB-derived PGS to predict into TPMI, or the TPMI-derived PGS when predicting into other cohorts, it is possible that PCs could confound the prediction. In summary, I recommend that you always adjust for sex, age, PCs in the PGS analyses. (Interestingly, I found that the r2 is also reported with covariates in the supplementary tables.) For the AUC, you can perhaps compare it to an AUC of a base model (with standard covariates).

Response: For explained variation (r^2), we followed the approach published by Ni et al.[36] and now report three different r^2 values. First, we report the raw r^2 without any covariate adjustment. Second, we report r^2 with standard covariates, including sex, age, and PCs. Finally, we include the partial r^2, which is derived from a comparison between the full model (including PRS) and the reduced model (excluding PRS), using the R package rsq. Additionally, because genetic PCs can

influence PRS-based predictions across cohorts, we now report r^2 with PCs included as covariates in the supplementary tables.

For AUC comparisons, we have followed the reviewer's recommendation by introducing a baseline model that includes only standard covariates (sex, age, and PCs). This approach allows for a clearer assessment of the added predictive value of PRS beyond conventional risk factors.

These revisions ensure that our PRS evaluation is robust, appropriately adjusted for relevant confounders, and methodologically transparent. We have updated the manuscript accordingly to reflect these improvements.

Revised sentences (in methods, LINE 974-983)

The explained variance (r^2) was used to evaluate the performance of PRS for quantitative traits[35,36], and two indices, area under the receiver operating characteristic curve (AUC) and liability-scaled $r^{2,}$ were used for PRS of dichotomized phenotypes. We followed the approach by Ni et al.[36] and report both raw r^2 and r^2 adjusted for covariates (sex, age, and PCs). Additionally, we include partial r^2 estimates, calculated using the R package rsq. To account for population stratification in cross-cohort predictions, we also report r^2 with PCs as covariates in the supplemental tables. For AUC comparisons, we include a baseline model incorporating standard covariates (sex, age, and PCs) to better assess the added predictive power of PRS.

12. Please provide logistic regression or linear regression GWAS summary statistics online, (as well as SAIGE or BOLT-LMM). This would make it much easier for researchers to use these in future studies, as many common methods have issues with using mixed model summary statistics, e.g. LDSC, LDpred2, PRS-CS, lassosum2, MEGA-PRS, SBayesS, etc. This would therefore also increase the impact of the study.

Response: To enhance the utility of our data, we will provide GWAS summary statistics from both logistic/linear regression (PLINK) and mixed models (e.g., SAIGE) in our publicly available resource, along with clear documentation for ease of use in future studies. We have noted this in the revised manuscript.

Revised sentences (in data availability, LINE 1043-1045)

All summary statistics, polygenic risk score (PRS) models, and GWAS results (summary statistics from SAIGE and PLINK) are available from the TPMI website (https://pheweb.ibms.sinica.edu.tw/

References

1 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015). https://doi.org/10.1038/nature15393

2 Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655-1664 (2009). https://doi.org/10.1101/gr.094052.109

3 Lin, L. Y., Warren-Gash, C., Smeeth, L. & Chen, P. C. Data resource profile: the National Health Insurance Research Database (NHIRD). *Epidemiol Health* **40**, e2018062 (2018). https://doi.org/10.4178/epih.e2018062

4 Schoeler, T. *et al.* Participation bias in the UK Biobank distorts genetic associations and downstream analyses. *Nat Hum Behav* **7**, 1216-1227 (2023). https://doi.org/10.1038/s41562-023-01579-9

5 Zhou, Y. *et al.* Performance of multigene testing in cytologically indeterminate thyroid nodules and molecular risk stratification. *PeerJ* **11**, e16054 (2023). https://doi.org/10.7717/peerj.16054

6 Tyagi, A., Goyal, A., Chaware, P. & Rathinam, B. A. D. Mutations of PHOX2B Gene in Patients of Obesity Hypoventilation Syndrome in Central India. *J Lab Physicians* **14**, 164-168 (2022). https://doi.org/10.1055/s-0041-1735582

7 He, D. *et al.* A longitudinal genome-wide association study of bone mineral density mean and variability in the UK Biobank. *Osteoporos Int* **34**, 1907-1916 (2023). https://doi.org/10.1007/s00198-023-06852-1

8 Wang, Y., Tsuo, K., Kanai, M., Neale, B. M. & Martin, A. R. Challenges and Opportunities for Developing More Generalizable Polygenic Risk Scores. *Annu Rev Biomed Data Sci* **5**, 293-320 (2022). https://doi.org/10.1146/annurev-biodatasci-111721-074830

9 Choi, S. W., Mak, T. S. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc* **15**, 2759-2772 (2020). https://doi.org/10.1038/s41596-020-0353-1

10 Wray, N. R. *et al.* Pitfalls of predicting complex traits from SNPs. *Nat Rev Genet* **14**, 507-515 (2013). https://doi.org/10.1038/nrg3457

11 Yang, H.-C. *et al.* The Taiwan Precision Medicine Initiative: A Cohort for Large-Scale Studies. *BIORXIV* (2024). https://doi.org/BIORXIV/2024/616932

12 Chen, C. Y. *et al.* Analysis across Taiwan Biobank, Biobank Japan, and UK Biobank identifies hundreds of novel loci for 36 quantitative traits. *Cell Genom* **3**, 100436 (2023). https://doi.org/10.1016/j.xgen.2023.100436

13 Kember, R. L. *et al.* Polygenic risk scores for cardiometabolic traits demonstrate importance of ancestry for predictive precision medicine. *Pac Symp Biocomput* **29**, 611-626 (2024).

14 Khunsriraksakul, C. *et al.* Multi-ancestry and multi-trait genome-wide association meta-analyses inform clinical risk prediction for systemic lupus erythematosus. *Nat Commun* **14**, 668 (2023). https://doi.org/10.1038/s41467-023-36306-5

15 Fritsche, L. G. *et al.* Cancer PRSweb: An Online Repository with Polygenic Risk Scores for Major Cancer Traits and Their Evaluation in Two Independent Biobanks. *Am J Hum Genet* **107**, 815-836 (2020). https://doi.org/10.1016/j.ajhg.2020.08.025

16 Mars, N. *et al.* Genome-wide risk prediction of common diseases across ancestries in one million people. *Cell Genom* **2**, None (2022). https://doi.org/10.1016/j.xgen.2022.100118

17 Mars, N. *et al.* Polygenic and clinical risk scores and their impact on age at onset and prediction of cardiometabolic diseases and common cancers. *Nat Med* **26**, 549-557 (2020). https://doi.org/10.1038/s41591-020-0800-0

18 Deutsch, A. J. *et al.* Type 2 Diabetes Polygenic Score Predicts the Risk of Glucocorticoid-Induced Hyperglycemia in Patients Without Diabetes. *Diabetes Care* **46**, 1541-1545 (2023). https://doi.org/10.2337/dc23-0353

19 Mars, N. *et al.* Systematic comparison of family history and polygenic risk across 24 common diseases. *Am J Hum Genet* **109**, 2152-2162 (2022). https://doi.org/10.1016/j.ajhg.2022.10.009
20 Ma, Y., Patil, S., Zhou, X., Mukherjee, B. & Fritsche, L. G. ExPRSweb: An online repository with polygenic risk scores for common health-related exposures. *Am J Hum Genet* **109**, 1742-1760 (2022). https://doi.org/10.1016/j.ajhg.2022.09.001
21 Sumpter, N. A. *et al.* Association of Gout Polygenic Risk Score With Age at Disease Onset and Tophaceous Disease in European and Polynesian Men With Gout. *Arthritis Rheumatol* **75**, 816-825 (2023). https://doi.org/10.1002/art.42393
22 Truong, B. *et al.* Integrative polygenic risk score improves the prediction accuracy of complex traits and diseases. *Cell Genom* **4**, 100523 (2024). https://doi.org/10.1016/j.xgen.2024.100523
23 Lai, D. *et al.* Evaluating risk for alcohol use disorder: Polygenic risk scores and family history. *Alcohol Clin Exp Res* **46**, 374-383 (2022). https://doi.org/10.1111/acer.14772
24 Namba, S. *et al.* Common Germline Risk Variants Impact Somatic Alterations and Clinical Features across Cancers. *Cancer Res* **83**, 20-27 (2023). https://doi.org/10.1158/0008-5472.CAN-22-1492
25 Tanigawa, Y. *et al.* Significant sparse polygenic risk scores across 813 traits in UK Biobank. *PLoS Genet* **18**, e1010105 (2022). https://doi.org/10.1371/journal.pgen.1010105
26 Jung, H. *et al.* Integration of risk factor polygenic risk score with disease polygenic risk score for disease prediction. *Commun Biol* **7**, 180 (2024). https://doi.org/10.1038/s42003-024-05874-7
27 Prive, F. *et al.* Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *Am J Hum Genet* **109**, 12-23 (2022). https://doi.org/10.1016/j.ajhg.2021.11.008
28 Chang, K. C. *et al.* Survey of hepatitis B virus infection status after 35 years of universal vaccination implementation in Taiwan. *Liver Int* **44**, 2054-2062 (2024). https://doi.org/10.1111/liv.15959
29 Bastarache, L. *et al.* The phenotype-genotype reference map: Improving biobank data science through replication. *Am J Hum Genet* **110**, 1522-1533 (2023). https://doi.org/10.1016/j.ajhg.2023.07.012
30 Bastarache, L. Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. *Annu Rev Biomed Data Sci* **4**, 1-19 (2021). https://doi.org/10.1146/annurev-biodatasci-122320-112352
31 Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* **31**, 1102-1110 (2013). https://doi.org/10.1038/nbt.2749
32 GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330 (2020). https://doi.org/10.1126/science.aaz1776
33 Taylor, D. J. *et al.* Sources of gene expression variation in a globally diverse human cohort. *Nature* **632**, 122-130 (2024). https://doi.org/10.1038/s41586-024-07708-2
34 Wang, Q. S. *et al.* The whole blood transcriptional regulation landscape in 465 COVID-19 infected samples from Japan COVID-19 Task Force. *Nat Commun* **13**, 4830 (2022). https://doi.org/10.1038/s41467-022-32276-2
35 Lee, S. H., Goddard, M. E., Wray, N. R. & Visscher, P. M. A better coefficient of determination for genetic profile analysis. *Genet Epidemiol* **36**, 214-224 (2012). https://doi.org/10.1002/gepi.21614
36 Ni, G. *et al.* A Comparison of Ten Polygenic Score Methods for Psychiatric Disorders Applied Across Multiple Cohorts. *Biol Psychiatry* **90**, 611-620 (2021). https://doi.org/10.1016/j.biopsych.2021.04.018

37 Chang, S. W. *et al.* A genome-wide association study on chronic HBV infection and its clinical progression in male Han-Taiwanese. *PLoS One* **9**, e99724 (2014). https://doi.org/10.1371/journal.pone.0099724
38 Zeng, Z. *et al.* Genome-wide association study identifies new loci associated with risk of HBV infection and disease progression. *BMC Med Genomics* **14**, 84 (2021). https://doi.org/10.1186/s12920-021-00907-0
39 Li, Y. *et al.* Genome-wide association study identifies 8p21.3 associated with persistent hepatitis B virus infection among Chinese. *Nat Commun* **7**, 11664 (2016). https://doi.org/10.1038/ncomms11664
40 Chen, H. H. *et al.* Host genetic effects in pneumonia. *Am J Hum Genet* **108**, 194-201 (2021). https://doi.org/10.1016/j.ajhg.2020.12.010
41 Covid- Host Genetics Initiative. A second update on mapping the human genetic architecture of COVID-19. *Nature* **621**, E7-E26 (2023). https://doi.org/10.1038/s41586-023-06355-3
42 Covid- Host Genetics Initiative. A first update on mapping the human genetic architecture of COVID-19. *Nature* **608**, E1-E10 (2022). https://doi.org/10.1038/s41586-022-04826-7
43 Asgari, S. & Pousaz, L. A. Human genetic variants identified that affect COVID susceptibility and severity. *Nature* **600**, 390-391 (2021). https://doi.org/10.1038/d41586-021-01773-7
44 Zhong, Y., Perera, M. A. & Gamazon, E. R. On Using Local Ancestry to Characterize the Genetic Architecture of Human Traits: Genetic Regulation of Gene Expression in Multiethnic or Admixed Populations. *Am J Hum Genet* **104**, 1097-1115 (2019). https://doi.org/10.1016/j.ajhg.2019.04.009
45 Gay, N. R. *et al.* Impact of admixture and ancestry on eQTL analysis and GWAS colocalization in GTEx. *Genome Biol* **21**, 233 (2020). https://doi.org/10.1186/s13059-020-02113-0
46 Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fine-mapping from summary data with the "Sum of Single Effects" model. *PLoS Genet* **18**, e1010299 (2022). https://doi.org/10.1371/journal.pgen.1010299
47 Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res* **51**, D977-D985 (2023). https://doi.org/10.1093/nar/gkac1010
48 Ojavee, S. E., Kutalik, Z. & Robinson, M. R. Liability-scale heritability estimation for biobank studies of low-prevalence disease. *Am J Hum Genet* **109**, 2009-2017 (2022). https://doi.org/10.1016/j.ajhg.2022.09.011

Referees' comments:

Referee #1 (Remarks to the Author):

The revised manuscript has thoroughly and comprehensively addressed all my concerns through a more nuanced exposition of the results highlighting ascertainment bias and reducing the over interpretation of the results. As such, I find the revised manuscript greatly improved and I support its publication.

We thank the reviewer for accepting our revised manuscript and endorsing its publication.

Referee #2 (Remarks to the Author):

I want to thank the authors for being responsive to feedback and improving their manuscript. I do not have any huge substantiative comments/concerns, but have listed some suggestions below to strengthen the manuscript, both in terms of framing and presentation.

We thank the reviewer for the helpful suggestions. We have further revised our paper accordingly.

Improving the narrative/flow of the piece:

1. The results would be strengthened by tying the TPMI results with the prevalence estimates either through NHIRD or through case numbers in UKBB. For example, in the section titled "cross-population comparison based on GWAS" (page 14, lines 534-542), it would have helped to tie the differential correlations with possible differences in the trait distributions. This would also influence the PRS scores, such as the higher performance of T2D (compared to SNP-heritability) compared to other traits.

   Per the reviewer's suggestion, we have revised the relevant sentences in the Results section and included the case/control numbers for both TPMI and UKB in Table S6 (The SNP-based heritability and cross-population genetic correlation).

   Revised sentences (in Results, line 536-546):

   Cross-population comparisons[1] with the Europeans from UKB showed varying degrees of genetic correlation, with strong, statistically significant correlations for traits like cholelithiasis ($\rho_{ge} > 0.999$), type 2 diabetes ($\rho_{ge} = 0.829$), and ischemic heart disease ($\rho_{ge} = 0.756$), but moderate correlations for gout ($\rho_{ge} = 0.616$) and psoriasis ($\rho_{ge} = 0.418$) (Table S6). The moderate correlations suggest the differentiated genetic mechanism and disease distribution across populations (gout case n = 24,411 in TPMI and 3,179 in UKB; psoriasis case n=4,166 in TPMI and 2,197 in UKB). Therefore, these findings demonstrate the importance of population-specific genetic studies,

as differences in genetic architectures between populations can significantly affect the accuracy of PRS models.

2. The possible selection biases with TPMI also ties directly in with the last section on overall health measures. IT makes sense that cardiometabolic diseases are enriched in a clinical and hospital setting. This will also influence the statistical power and predictive ability of genetics. I would have liked to see this explicitly explored/mentioned, either in the results or discussion (or both to tie it all in throughout).

   We have revised the results and discussion section as follows:

   Revised results sentences (Line 633-638):

   Among the identified clusters, the cardiometabolic disease cluster contributed the most to the indices, accounting for 1.32% of clinical visits (p-value = 0.02) and 3.55% of hospitalizations (p-value = 7.10×10^{-9}). This may reflect the high prevalence in the TPMI cohort recruited from the hospitals and long-term medical needs associated with cardiometabolic diseases.

   Revised discussion sentences (Line 720-727):

   By integrating these well-developed PRS models, we estimate that genetics account for 10.3% of variation of hospitalization duration in TPMI. Although the estimates of genetic contributions to health measure may be influenced by disease prevalences and ascertainment biases, including participation bias and time-related biases such as right and left censoring, our result suggest integrating genetic risk-based health management strategies with traditional risk factors, such as age, sex, smoking status, and BMI, may enhance prediction models and refine personalized risk stratification.

3. This is not necessary for publication, but I think it would be great to have at least one vignette throughout, such as HBV. So describe the prevalence, how it compares to population-level estimates. Then to discovery, genetic correlations, and PGS. And finally the contribution to overall health. That would make the piece seem more cohesive and give readers an example to point to. Right now there is so much (which is great!) but so much gets buried in the figures/supplement. A narrative is always appreciated! You do have this at the end for the discussion, but having all the results laid out, maybe in a box, would help set up this discussion better.

   Per the reviewer's insightful suggestion, we have added a box highlighting our findings in HBV.

Box 1. TPMI Findings on Hepatitis B
Hepatitis B is an endemic disease in Taiwan, with a 5-year prevalence of 2.7% according to the NHIRD. In TPMI, we identified 23,618 cases from linked EMR, whereas only 132 cases were identified among UKB Europeans. This large sample size enhances the power of GWAS findings. Among the 26 fine-mapped signals, 19 have not been previously reported as associated with hepatitis B in the GWAS Catalog. However, many of these (18) are correlated with liver function or related phenotypes. A phenome-wide pairwise genetic correlation analysis further revealed a significant negative correlation between hepatitis B and other autoimmune diseases, such as Sicca syndrome, psoriasis, and systemic lupus erythematosus. These previously undisclosed genetic findings also enabled the development of a well-performing polygenic risk model, validated across various biobanks. Our findings highlight the importance of conducting genetic studies in diverse populations, not only to identify population-specific causal variants but also to improve our understanding of diseases that disproportionately affect certain populations.

Minor presentation/word-choice comments:

Figure 1: Would have liked to see the distribution of correlations by trait type. For example, it looks like there are a cluster of respiratory traits that have a much lower prevalence in TPMI compared to the prevalence in NHIRD. This would make sense for something like asthma. This would also help give an idea of what types of outcomes are enriched in TPMI compared to NHIRD as those above the identity line. This further informs your genetics results, such as the lower replication for respiratory traits (which you do attribute to limited case numbers, but mentioning this further up explicitly would strengthen.)

Figure 2B would be stronger as a bar chart with the trait names above it.

My eyesight isn't great, so take this with a slight grain of salt, but please have mercy on my aging eyes and make all of the text in figures much larger/clearer. I would even make the lines/points larger. Even with zooming in, it was a challenge!

We revised our plots following the reviewer's suggestion by changing Fig. 1B to a bar chart and enlarging the text and data points. Additionally, we included scatter plots of the TPMI case proportions and NHIRD prevalence by disease category as Extended Data Fig. 1. We also enlarged the text for other figures, please find them in the revised manuscript.

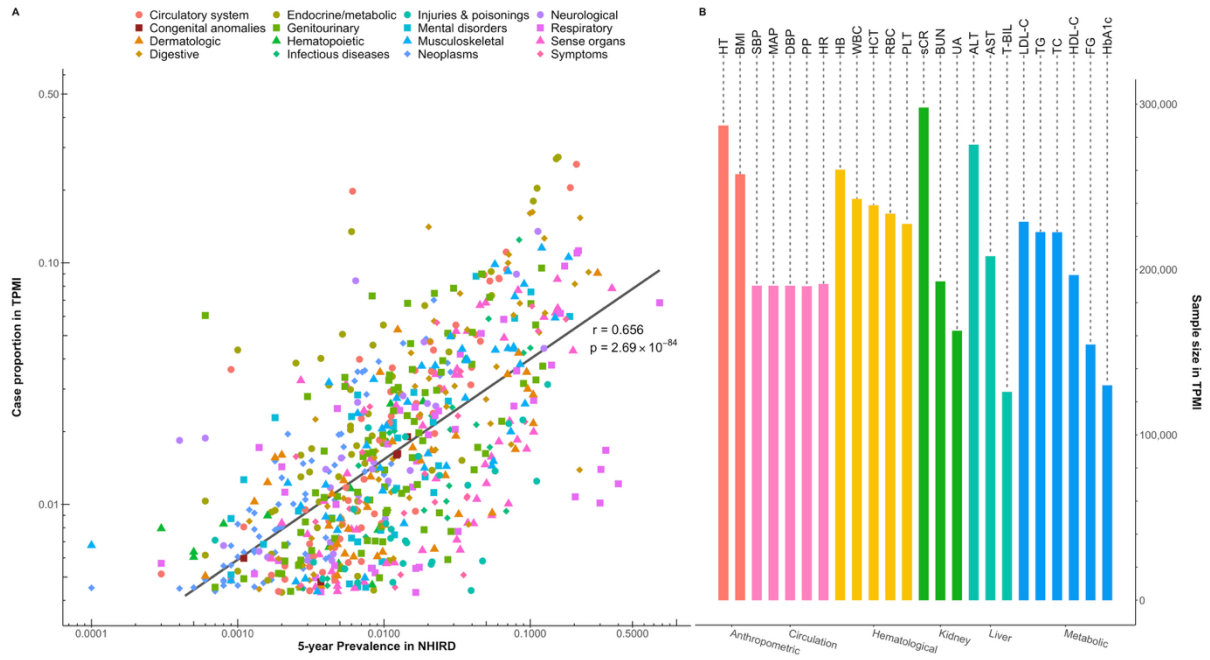


Figure R2-1. (The revised Fig. 1) The scatter plot of TPMI case proportion for phecodes and the bar chart of sample size for quantitative traits.

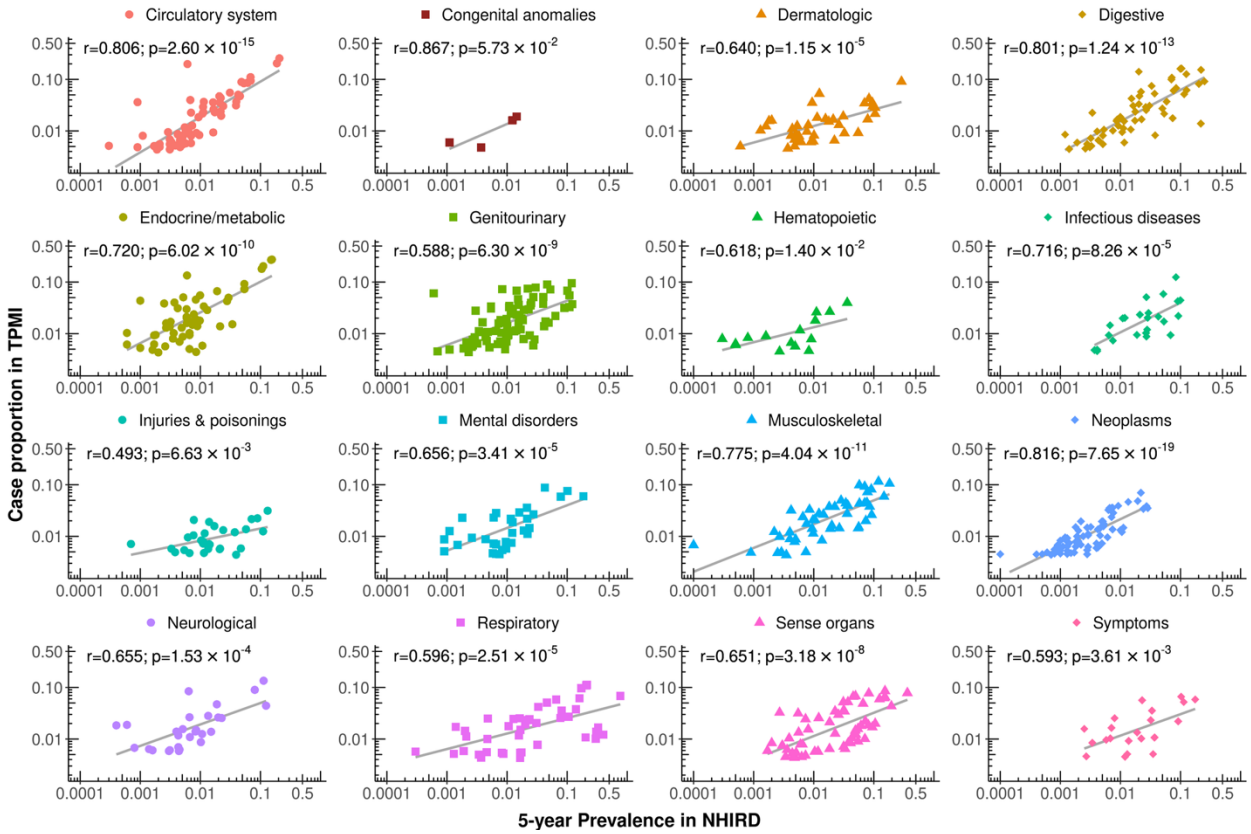


Figure R2-2. (The revised Extended Fig. 1) Scatter plots of the case proportion for dichotomized phenotypes by disease categories.

Line 676 calls HBV a population-unique disease. It is not unique to this population, but is enriched. Perhaps population-enriched instead of unique?

We now used 'population-enriched' instead of population-unique disease

Revised sentences (in Discussion, line 680-683):

Our unexpected success of GWAS and PRS for hepatitis B not only demonstrate the power of the large sample size of TPMI, but also reveal the necessity of population-specific genetic study for population-enriched diseases.

Line 715-716 states that genetics explains 10.3% of variation in hospital duration in Taiwan. It is more specifically in TPMI. Given the selection biases presented in the results, I would be cautious about extrapolating to the entire country.

We narrowed our claim per the reviewer's suggestion.
Revised sentences (in Discussion, line 720-722):

By integrating these well-developed PRS models, we estimate that genetics account for 10.3% of variation of hospitalization duration in TPMI.

Referee #3 (Remarks to the Author):

I would like to thank the authors for a very through response. I think the paper has improved substantially and very interesting! I found the comparison of hertiabilities between biobanks to be interesting. I'm stubbornly still surprised by this. Could this be due to an overestimated effective sample size in the LDSC, which would lead to downward bias? (probably not) Perhaps use a lower relatedness threshold? See perhaps also https://pmc.ncbi.nlm.nih.gov/articles/PMC10066905/ (which describes a related issue, but I don't think this is a plausible explanation here). Lastly, perhaps you can shed some light on this by expanding the LDpred2/SBayesS heritability estimates to more common outcomes. Those methods may have advantages over LDSC when it comes to dealing with varying SNP sample size issues. It would at least be interesting to see whether they are very different for, e.g. height and BMI. Another possible explanation is that height and BMI are subject to less assortative mating biases than in, e.g. the UK. (See e.g. Yengo et al. 2018, and others)

The discussion on accurately estimating heritability has been long-standing. As the reviewer noted, many factors can affect heritability estimation, including assumptions underlying statistical approaches, the phenotyping process, the degree of relatedness among participants, and the genetic variants used. We agree with the reviewer's point. Since this discussion focuses on quantitative traits, specifically body height and BMI, issues such as overestimated effective sample size and liability-scale transformation should not be of

concern. We also examined the effect of using different relatedness thresholds for defining unrelated individuals. Previously, we used third-degree relatedness ($\hat{p}<0.09375$) as the cutoff, resulting in an unrelated set of n = 248,754. We have now applied stricter thresholds of $\hat{p}<0.0442$ and $\hat{p}<0.0221$ resulting in sample sizes of n = 237,987 and n = 179,977, respectively. The results indicate no significant differences across these thresholds. (body height: h^2=0.323±0.013 for $\hat{p}<0.09375$, 0.334±0.015 for $\hat{p}<0.0442$, and 0.346±0.018 for $\hat{p}<0.0221$; BMI: h2=0.218±0.009 for $\hat{p}<0.09375$, 0.220±0.011 for $\hat{p}<0.0442$, and 0.218±0.012 for $\hat{p}<0.0221$).

In our analysis, we observed comparable heritability estimates between TPMI and the UK Biobank (Table S6). To address comments raised by all three reviewers, we further explored the literature and found our estimates are in close agreement with those reported for the Taiwan Biobank, Biobank Japan, and UK Biobank, derived with the same estimation approaches we used (Figure R3-1 and Table R3-1)[2]. Following the previous reviewer's suggestion, we also performed different software tools for heritability estimation, including LDPred2 and SBayesS. We found high concordance between LDSC and LDPred2 estimates, while SBayesS systemically yielded lower heritability estimates (Figure R3-2 and Table R3-1). This trend was also observed for both body height and BMI. We acknowledge that these differences likely reflect varying statistical assumptions. A previous simulation study suggests that LDSC estimates are robust to confounding from shared environmental factors and provide a reasonable lower-bound estimate of heritability based on common variants.[3]

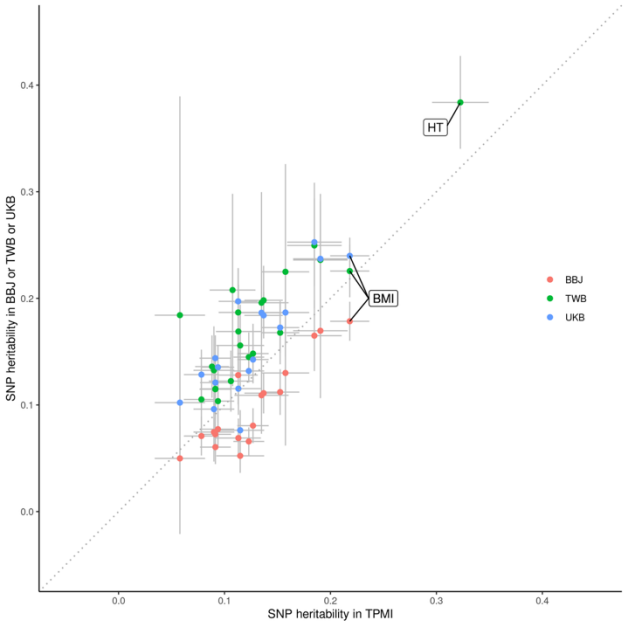


Figure R3-1. The estimates of heritability for 24 quantitative traits in TPMI (x-axis) and other biobanks (y-axis), including Taiwan Biobank (TWB, green), Biobank Japan (red), and UK Biobank (blue). The bars indicate the 95% confidence interval.

Moreover, our estimates do not account for the contribution of rare variants (MAF < 0.01). In a study of 25,465 unrelated Europeans from TOPMed project, common SNP-based heritability estimates were 0.48 ± 0.02 for height and 0.24 ± 0.02 for BMI using the GREML (residual maximum likelihood) approach[4], which generally reports higher heritability compared to LDSC[3]. When sequencing data were used, the heritability estimates increased

to 0.68 ± 0.10 for height and 0.30 ± 0.10 for BMI, indicating the significant contribution of rare variants and a limitation of our current dataset [4].

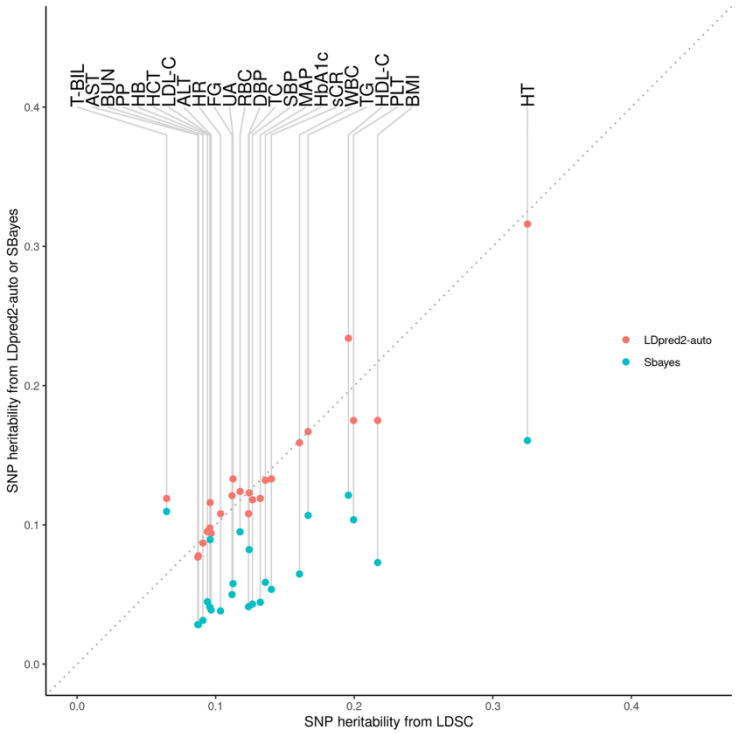


Figure R3-2. The scatter plot of comparison of heritability estimation between LDSC (x-axis) and LDpred2-auto (red) and SBayesS (blue).

We thank the reviewer for raising the issue of assortative mating. Prior studies in Europeans (e.g., UK Biobank[5]) show that assortative mating can inflate heritability estimates for height by 14–23%. Evidence in Chinese populations is more limited; Li et al. reported a modest spousal correlation for height (r = 0.141) among Han Chinese, and studies for BMI suggest weak or non-significant spousal correlations[5]. Applying the method of Yengo et al. in TPMI, we found no significant correlation between odd- and even-chromosome PRS for height (θ = 0.013, p = 0.067) or BMI (θ = 0.010, p = 0.155), unlike the stronger effect seen in Europeans (θ = 0.032, p = 6.5×10^{-89})[6]. These results suggest minimal assortative mating bias in our heritability estimates. We have clarified this in the Discussion and emphasized the need for further research in East Asian cohorts.

Revised sentences (in Discussion, line 751-755):

Additionally, we observed a relatively low estimated heritability for body height and BMI in TPMI, compared to values reported in the literature.[2,4] These estimates may be affected by factors such as inconsistencies in assessment across hospitals in TPMI cohort, variations in statistical approaches, and reduced bias of assortative mating for height in Han Chinese population.[5-7]

Table R3-1. The estimated heritability of 24 quantitative traits accoss biobanks and approaches

	TPMI					Taiwan Biobank		Biobank Japan		UK Biobank	
Trait	LDSC		LDPred2	SbayesS		LDSC		LDSC		LDSC	
ALT	0.091	±0.007	0.094	0.039	±0.002	0.115	±0.014	0.061	±0.008	0.121	±0.010
AST	0.078	±0.008	0.077	0.028	±0.002	0.105	±0.016	0.071	±0.009	0.129	±0.012
T-BIL	0.058	±0.012	0.119	0.110	±0.003	0.184	±0.105	0.050	±0.016	0.102	±0.024
BMI	0.218	±0.009	0.175	0.073	±0.003	0.226	±0.013	0.179	±0.009	0.240	±0.009
BUN	0.088	±0.008	0.078	0.028	±0.002	0.136	±0.015				
DBP	0.123	±0.007	0.108	0.041	±0.002	0.145	±0.012	0.066	±0.007	0.132	±0.006
FG	0.115	±0.012	0.121	0.050	±0.003	0.156	±0.020	0.052	±0.008	0.076	±0.010
HbA1c	0.135	±0.013	0.132	0.059	±0.003	0.196	±0.053	0.109	±0.018	0.187	±0.013
HB	0.091	±0.008	0.095	0.045	±0.002	0.115	±0.012	0.073	±0.008	0.144	±0.011
HCT	0.094	±0.008	0.098	0.041	±0.002	0.104	±0.010	0.077	±0.009	0.135	±0.010
HDL-C	0.191	±0.013	0.234	0.121	±0.003	0.236	±0.032	0.170	±0.032	0.237	±0.028
HT	0.323	±0.014	0.316	0.161	±0.004	0.384	±0.022				
HR	0.106	±0.011	0.108	0.038	±0.002	0.122	±0.015				
LDL-C	0.090	±0.010	0.116	0.089	±0.002	0.133	±0.021	0.0746	±0.0141	0.0961	±0.0124
MAP	0.132	±0.008	0.119	0.044	±0.002						
PLT	0.185	±0.013	0.175	0.104	±0.003	0.250	±0.030	0.165	0.017	0.253	±0.021
PP	0.092	±0.006	0.087	0.031	±0.002						
RBC	0.113	±0.009	0.124	0.095	±0.002	0.187	±0.021	0.128	±0.014	0.197	±0.015
SBP	0.127	±0.008	0.118	0.043	±0.002	0.148	±0.014	0.081	±0.008	0.143	±0.006
sCR	0.137	±0.009	0.133	0.054	±0.002						
TC	0.113	±0.011	0.123	0.082	±0.002	0.169	±0.022	0.069	±0.009	0.115	±0.013
TG	0.158	±0.011	0.167	0.107	±0.003	0.225	±0.052	0.130	±0.035	0.187	±0.021
UA	0.108	±0.011	0.133	0.058	±0.002	0.208	±0.046				
WBC	0.153	±0.009	0.159	0.065	±0.003	0.168	±0.015	0.112	±0.011	0.173	±0.012

References

1 Brown, B. C., Asian Genetic Epidemiology Network Type 2 Diabetes, C., Ye, C. J., Price, A. L. & Zaitlen, N. Transethnic Genetic-Correlation Estimates from Summary Statistics. *Am J Hum Genet* **99**, 76-88 (2016). https://doi.org/10.1016/j.ajhg.2016.05.001

2 Chen, C.-Y. *et al*. Analysis across Taiwan Biobank, Biobank Japan, and UK Biobank identifies hundreds of novel loci for 36 quantitative traits. *Cell Genomics* **3**, 100436 (2023). https://doi.org/https://doi.org/10.1016/j.xgen.2023.100436

3 Evans, L. M. *et al*. Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nat Genet* **50**, 737-745 (2018). https://doi.org/10.1038/s41588-018-0108-x

4 Wainschtein, P. *et al*. Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. *Nat Genet* **54**, 263-273 (2022). https://doi.org/10.1038/s41588-021-00997-7

5 Li, M. X. *et al*. A major gene model of adult height is suggested in Chinese. *J Hum Genet* **49**, 148-153 (2004). https://doi.org/10.1007/s10038-004-0125-8

6 Yengo, L. *et al*. Imprint of assortative mating on the human genome. *Nat Hum Behav* **2**, 948-954 (2018). https://doi.org/10.1038/s41562-018-0476-3

7 Border, R. *et al*. Assortative mating biases marker-based heritability estimators. *Nat Commun* **13**, 660 (2022). https://doi.org/10.1038/s41467-022-28294-9

Referees' comments:

Referee #3 (Remarks to the Author):

I would like to thank the authors for exploring the heritability estimation further. I think this is really interesting analysis. I would recommend the authors to include some of the analysis in the main manuscript or supplementary analysis, as it is in my opinion a very interesting and important finding. Also, the paper cited to argue the choice of LDSC, predates and does not cite LDpred2 or SBayesS. The benefit from using LDpred2 and SBayesS is that it estimates the heritability under a non-infinitesimal prior, meaning that the estimates should in theory be more accurate for outcomes that do not have an infinitesimal genetic architecture.

We thank the reviwer for his/her suggestion and we now have included our findings and disscussions regarding the heriatibiluty estimation in the supplementary material.
Here is the new section about the heritability in the supplmentary matrial:

Supplementary text P.27-28

SNP-based heritability for quantitative traits

We performed the linkage disequilibrium score regression (LDSC) for estimating the SNP-based heritability for all 24 quantitative traits[1]. The GWAS summary statistics from the linear regression with TPMI unrelated set ($\hat{p}$<0.09375 [3rd degree of relatedness], n = 248,754) were used. The estimated heritability from TPMI ranged from 0.088±0.008 for BUN to 0.323±0.014 for HT. To assess the robustness of our estimates, we compared our results with reported heritability values from the Taiwan Biobank (TWB), Biobank Japan (BBJ), and UK Biobank (UKB)[2]. We observed strong and statistically significant correlations between TPMI and the other three biobanks (r = 0.88 and p-value = 1.05×10^{-5} with TWB; r = 0.90 and p-value = 1.12×10^{-6} with BBJ; and r = 0.84 and p-value = 2.56×10^{-5} with UKB, Fig. 19 and Table 4). However, the heritability estimate for HT from TPMI (h^2 = 0.323 ± 0.014) was slightly lower than that from TWB (h^2 = 0.384

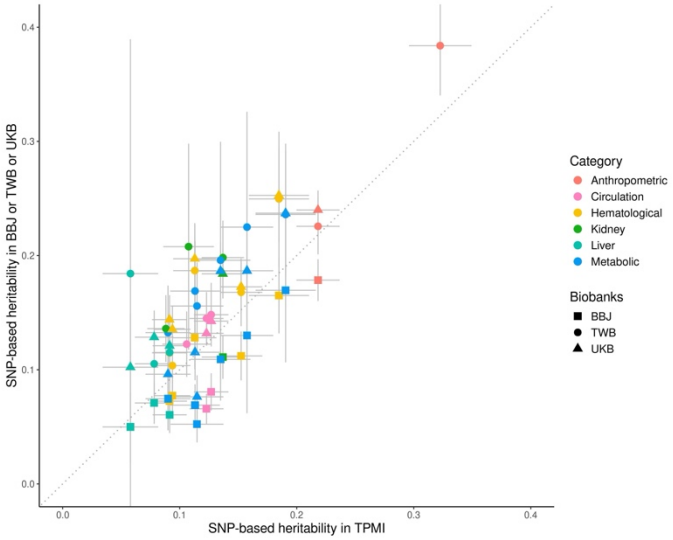


Figure 19. The estimates of heritability for 24 quantitative traits in TPMI (x-axis) and other biobanks (y-axis), including Taiwan Biobank (TWB, circle), Biobank Japan (BBJ, square), and UK Biobank (triangle). The bars indicate the 95% confidence interval.

± 0.022; Fig. 19 and Table 4). This may be explained by the fact that the EMR-extracted body height includes both self-reported and measured values, which may introduce greater phenotypic variation compared to the Taiwan Biobank, where measurements were conducted by trained personnel using a standardized protocol.

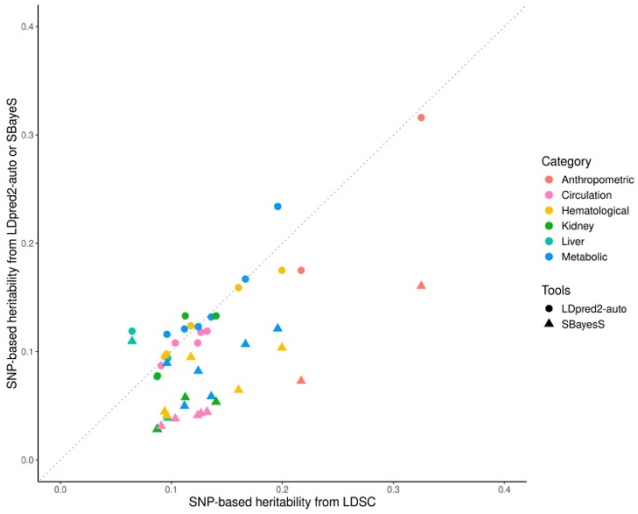


Figure 20. The scatter plot of comparison of heritability estimation between LDSC (x-axis) and LDpred2-auto (circle) and SBayesS (triangle).

We then further investigated the effect of parameter settings and statistical approaches on heritability estimation. First, we applied stricter thresholds for defining the unrelated set ($\hat{p}$<0.0442 and $\hat{p}$<0.0221), resulting in sample sizes of n = 237,987 and n = 179,977, respectively. The results indicated no significant differences across these thresholds. (HT: h^2=0.323±0.013 for $\hat{p}$<0.09375, 0.334±0.015 for $\hat{p}$<0.0442, and 0.346±0.018 for $\hat{p}$<0.0221; BMI: h2=0.218±0.009 for $\hat{p}$<0.09375, 0.220±0.011 for $\hat{p}$<0.0442, and 0.218±0.012 for $\hat{p}$<0.0221). Next, we used different software tools for heritability estimation, including LDPred2-auto[3] and SBayesS[4], which incorporate a non-infinitesimal prior. We found high concordance between the estimates from LDSC and LDPred2-auto, while SBayesS consistently produced lower heritability estimates (Figure 20 and Table 4).

Lastly, differences in the degree of assortative mating across populations may also impact heritability estimation. A previous study in the UK Biobank showed that assortative mating can inflate heritability estimates for height by 14–23%[5]. However, among Han Chinese, a modest spousal correlation for height (r = 0.141) and a non-significant spousal correlation for BMI have been reported[5]. We also applied the method of Yengo et al. in TPMI, and we found no significant correlation between odd- and even-chromosome PRS for height (θ = 0.013, p = 0.067) or BMI (θ = 0.010, p = 0.155), in contrast to the stronger effect observed in Europeans (θ = 0.032, p = 6.5×10^{-89})[6]. These results suggest minimal assortative mating bias in our heritability estimates.

Table 4. The estimated heritability of 24 quantitative traits accoss biobanks and approaches

	TPMI					Taiwan Biobank[2]		Biobank Japan[2]		UK Biobank[2]	
Trait	LDSC		LDPred2	SbayesS		LDSC		LDSC		LDSC	
ALT	0.091	±0.007	0.094	0.039	±0.002	0.115	±0.014	0.061	±0.008	0.121	±0.010
AST	0.078	±0.008	0.077	0.028	±0.002	0.105	±0.016	0.071	±0.009	0.129	±0.012
T-BIL	0.058	±0.012	0.119	0.110	±0.003	0.184	±0.105	0.050	±0.016	0.102	±0.024
BMI	0.218	±0.009	0.175	0.073	±0.003	0.226	±0.013	0.179	±0.009	0.240	±0.009
BUN	0.088	±0.008	0.078	0.028	±0.002	0.136	±0.015				
DBP	0.123	±0.007	0.108	0.041	±0.002	0.145	±0.012	0.066	±0.007	0.132	±0.006
FG	0.115	±0.012	0.121	0.050	±0.003	0.156	±0.020	0.052	±0.008	0.076	±0.010
HbA1c	0.135	±0.013	0.132	0.059	±0.003	0.196	±0.053	0.109	±0.018	0.187	±0.013
HB	0.091	±0.008	0.095	0.045	±0.002	0.115	±0.012	0.073	±0.008	0.144	±0.011
HCT	0.094	±0.008	0.098	0.041	±0.002	0.104	±0.010	0.077	±0.009	0.135	±0.010
HDL-C	0.191	±0.013	0.234	0.121	±0.003	0.236	±0.032	0.170	±0.032	0.237	±0.028
HT	0.323	±0.014	0.316	0.161	±0.004	0.384	±0.022				
HR	0.106	±0.011	0.108	0.038	±0.002	0.122	±0.015				
LDL-C	0.090	±0.010	0.116	0.089	±0.002	0.133	±0.021	0.0746	±0.0141	0.0961	±0.0124
MAP	0.132	±0.008	0.119	0.044	±0.002						
PLT	0.185	±0.013	0.175	0.104	±0.003	0.250	±0.030	0.165	0.017	0.253	±0.021
PP	0.092	±0.006	0.087	0.031	±0.002						
RBC	0.113	±0.009	0.124	0.095	±0.002	0.187	±0.021	0.128	±0.014	0.197	±0.015
SBP	0.127	±0.008	0.118	0.043	±0.002	0.148	±0.014	0.081	±0.008	0.143	±0.006
sCR	0.137	±0.009	0.133	0.054	±0.002						
TC	0.113	±0.011	0.123	0.082	±0.002	0.169	±0.022	0.069	±0.009	0.115	±0.013
TG	0.158	±0.011	0.167	0.107	±0.003	0.225	±0.052	0.130	±0.035	0.187	±0.021
UA	0.108	±0.011	0.133	0.058	±0.002	0.208	±0.046				
WBC	0.153	±0.009	0.159	0.065	±0.003	0.168	±0.015	0.112	±0.011	0.173	±0.012

1 Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-295 (2015). https://doi.org/10.1038/ng.3211

2 Chen, C.-Y. *et al.* Analysis across Taiwan Biobank, Biobank Japan, and UK Biobank identifies hundreds of novel loci for 36 quantitative traits. *Cell Genomics* **3**, 100436 (2023). https://doi.org/https://doi.org/10.1016/j.xgen.2023.100436

3 Prive, F., Albinana, C., Arbel, J., Pasaniuc, B. & Vilhjalmsson, B. J. Inferring disease architecture and predictive ability with LDpred2-auto. *Am J Hum Genet* **110**, 2042-2055 (2023). https://doi.org/10.1016/j.ajhg.2023.10.010

4 Zeng, J. *et al.* Widespread signatures of natural selection across human complex traits and functional genomic categories. *Nat Commun* **12**, 1164 (2021). https://doi.org/10.1038/s41467-021-21446-3

5 Li, M. X. *et al.* A major gene model of adult height is suggested in Chinese. *J Hum Genet* **49**, 148-153 (2004). https://doi.org/10.1007/s10038-004-0125-8

6 Yengo, L. *et al.* Imprint of assortative mating on the human genome. *Nat Hum Behav* **2**, 948-954 (2018). https://doi.org/10.1038/s41562-018-0476-3