nature portfolio

Peer Review File

Discovery and prioritization of genetic determinants of kidney function in 297,355 individuals from Taiwan and Japan.

Corresponding Author: Dr Chin-Chi Kuo

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

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
In “Discovery and prioritization of genetic determinants of kidney function in approximately 310,000 individuals from ESKD-prevalent Taiwan and Japan”, Chen et al conduct a GWAS of eGFR in Japanese and Taiwanese cohorts. They discover significant SNPs, conduct conditional analyses, conduct eQTL studies with publicly available data, conduct geneset enrichment with public available data, evaluate overlap with regulatory data, and create and evaluate a polygenic risk score. This paper follows a similar approach to many other GWAS of eGFR and CKD that have been published in the past five years. The authors state that the novel contribution here is the fact that this study is being done in a relatively large East Asian cohort, which has been under-represented historically and that this new population leads to new discoveries unable to be found in other populations.
Overall, this study does not seem to illuminate any major new findings. Moreover, there are a number of unconventional choices, some errors, and debatable assertions that raise doubts about the rigor and depth overall of the study. The title is illustrative of this. It is unclear what “ESKD-prevalent” means or its significance and highlighting an approximate number of samples is also unclearly motivated. The choice to seemingly report raw numbers of SNPs that were significant or replicated (e.g. lines 129 and 141) is confusing. There is no real depth of pursuit of the majority of lists of genes reported. The Bonferroni correction on lines 153-154 is incorrect and it is concerning that the authors didn’t pick up on the fact that the corrected p-value they report was almost genome-wide significant for far fewer tests. Performing targeted coloc with only significant SNPs rather than a full coloc between the GWAS and the eQTL data is debatable. And it also doesn’t have face validity that the “The most intriguing discovery is that the majority of SNPs associated with decreased kidney function in East Asian populations differ from those associated with decreased kidney function in Caucasian populations.” There are many other reasons for this observation which have profoundly different implications. Finally, the data availability statement is not in line with most work published in this era. At the very least, GWAS summary statistics and coloc statistics should be openly shared. Individual level data with disease age of onset and eGFR should also be able to be shared to a federated public resource without compromising patient privacy.

Reviewer #2

(Remarks to the Author)
The manuscript presented by Chen at al. reports on a very important study. The study has been conducted professionally and up to current standards. The manuscript is very well-written. It is of high relevance and importance to the field.

I do not have major concerns, but only a few minor remarks.

* Data availability: If possible, it would be beneficial to provide access to summary-level meta-analysis results of eGFR and BUN, this should not threaten privacy of participants

* Kidney cell types: It could potentially be interesting to use LDsc-seg with kidney cell type data sets. You already applied this approach to investigate regulatory elements.
This might help prioritize genes specifically expressed in certain kidney cell types.

* Figure 2: Luckily, there are no green dots, please adjust caption.
* Table 1: Is the last pathway significant (p-adj > 0.05) or should it be removed?
* Methods, line 456 - isn't this b37 => b38?

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
Thank you for revising your manuscript. I appreciate your detailed responses.

Reviewer #3

(Remarks to the Author)
In this study, Chen et al conducted a meta-analysis of GWAS for kidney function based on the existing databases including Taiwan and Japan biobank. Although the study merits from a large sample size and the East Asian population, the manuscript contains a number of major issues. Many technical words in the manuscripts are non-standard in the field of human genetic epidemiology. The English writing is too problematic.

Abstract:
- “ancestral diversity”:
- the first sentence in the abstract is very strange.
- “238 independent signals”: statistical independent?
- line 43-45: grammar problem
- we do not say “substantially differentiate” (line 46) to express classification ability of a predictor
Introduction:
- line 55, “definitive cure”: not standard
- line 61, “the identification of genetic variants that cause CKD is crucial for accelerating drug development”: from this study, you can at most find susceptible loci rather than finding the causal loci for CKD.
- Line 63, “More than 250 genetic loci associated with biomarkers of CKD”: biomarker is a very broad term that in genetic association studies it is usually reserved to indicate variants rather than phenotype
- Line 65, “reproducible”: not an appropriate word here.
Since almost every sentence has some problem, I stopped correcting them from here.
Results:
- It is not very often to see that a meta-analysis discovers over 5000 genome-wide significant markers. As a standard way of presentation, the authors should output a table in the main result (not supplement) including: the summary statistic (effect size, p-value, maybe MAF) of the top variants in each dataset, and the meta-analysis summary statistic. The readers will need to check the consistency of the effect size, and interpret the increased power that generates a smaller p-value.
- The PRS includes age and gender variable. I wonder that apart from the age, how well can the PRS based on pure genetic information can differentiate the high risk and low risk population. The study also analyzed that the heritability estimated for eGFR is only 10.9%. The PRS is constructed based on the eGRF variants to predict the CKD outcome, so I would expect that the predictability of genetic risk factor is not very strong but the power of PRS is mainly contributed by age. If that is the case, the suggested PRS as a major conclusion of the study may not be very specific for CKD or eGFR.
Methods:
- Use of CMUH data is unclear. The paragraph only described the clinical variables in the database, but did not include the genetic data, yet the paragraph is titled “genetic data from CMUH”. Does the author mean that the Taiwan biobank can matched to the subject in this database? In that case, you should say “match” rather than “integrate” in line 434. The ultimate source of the “genetic data from CMUH” is still unclear. Population included in this data should also be described.
- BBJ: Population in this data should be directly described. This information is guessed from the mentioning of “University of Tokyo”.
- Phenotype identification: the summary statistic of the phenotype variables, eGFR and BUN, should be provided.
- Line 498, any reference to support the cutoff of LD = 0.6 to define “independent SNPs”. It may still not accurate to claim them as “independent” unless they are statistically independent (by definition).
- Line 499: this sentence is logically not correct.
- Any reference to support the “merging” of “loci”? not a standard treatment
- Line 519: defining a marker as “replicated” in a second cohort by p-value < 0.05 is debatable. Better to directly report the p-values in the replication dataset.
- Line 525: please explain why negative effect on BUN is considered as relevant to kidney function. Strong biological assumption is made here on SNP effect?

Version 2:

Reviewer comments:

Reviewer #3

(Remarks to the Author)
1) The English of the manuscript is still not okay. There are problems in nearly every 2 sentences. For instance,
Line 53, “An 8-year difference between high and low genetic risk groups was observed from the age of 50.” – not sure what the author means.
Line 60, “Dapagliflozin is the only oral drug approved by the U.S. Food and Drug Administration to slow the progression of CKD, and this therapeutic effect was an unexpected discovery.”
2) Scientifically, the manuscript lacks novelty. It is a meta-analysis study, without new data nor new method.
3) Results: I’m also concerned with the number of reported genome-wide significant markers - 5,790 is too large. In the largest meta-analysis GWAS for Alzheimer’s disease with heritability of nearly 80%, sample size is >400,000, and the number of significant markers is 29.
In a similar meta-analysis study involving UK Biobank and Finnish cohort on Pneumonia with a total sample size > 500,000, the number of independent genome-wide significant biomarker is 2, and the gene-based analysis identifies 18 genes reaching the significance level. (Campos AI, Kho P, Vazquez-Prada KX, García-Marín LM, Martin NG, Cuéllar-Partida G, Rentería ME. Genetic susceptibility to pneumonia: a GWAS meta-analysis between the UK Biobank and FinnGen. Twin Research and Human Genetics. 2021 Jun;24(3):145-54.)
Since the heritability of kidney disease is 30-75%, we would expect that with a sample size is ~250,000, the number of significant markers would also be in the range of 1-50. But the author reported 5,790! As the author used standard software, I cannot tell from the writing where went wrong in their analysis. But >5000 significant markers do not seem right.
4) Conclusion is too strong:
Line 54, “In countries with a high incidence of ESKD, PRSCKD can be used for early kidney disease prevention.” – this conclusion is too strong at this stage.

Reviewer #4

(Remarks to the Author)
This is not a comprehensive review of the manuscript, but I would like to provide some specific feedback:

In the chapter "Genetic correlations of the eGFR and BUN with other phenotypes," there is an inconsistency in the reported p-value for the correlation between BUN and S-Cre. The manuscript lists it as 4.06x10^-10, whereas Table ST8 lists it as 4.06x10^-8. Please correct this discrepancy in the manuscript or the table. Please also see Supplementary Figure 3b for this.

In the abstract, the sentence "An 8-year difference between high and low genetic risk groups was observed from the age of 50" is unclear. I recommend removing this.

In the introduction, line 71, the phrase "common variations" should be revised to "common variants." Here, a "variant" refers to genetic differences such as SNPs, indels, or deletions.

In the discussion, last paragraph, line 396, the mention of "2,140 novel SNPs" appears without prior context. I could not find any reference to this number elsewhere in the manuscript. Please clarify where this figure originates and provide the relevant details in the Results section, or remove this from the discussion.

Version 3:

Reviewer comments:

Reviewer #4

(Remarks to the Author)
I thank the authors for addressing all my comments.

Reviewer: 1

Comments to the Author

1. ***Overall, this study does not seem to illuminate any major new findings. Moreover, there are a number of unconventional choices, some errors, and debatable assertions that raise doubts about the rigor and depth overall of the study. The title is illustrative of this. It is unclear what "ESKD-prevalent" means or its significance and highlighting an approximate number of samples is also unclearly motivated.***

Response:

We thank the reviewer for his/her comments. Taiwan and Japan have the highest prevalence of end-stage kidney disease (ESKD) worldwide, at 3,679 per million population (pmp) and 2,696 pmp, respectively. This makes our study particularly unique, as these populations may possess a genetic predisposition to impaired kidney function.[1] However, we agree with the reviewer that this vulnerability may not solely stem from genetic factors.

On the other hand, our work could contribute significantly to CKD genetics research, especially for the East-Asian ethnicity group, which, despite its high ESKD prevalence, has historically been under-represented in such studies. The present study involves a total of 297,355 patients of East-Asian ancestry and is the largest East-Asian study to date, according to the GWAS Catalog.[2] Up to date, the most comprehensive genetic analysis of CKD involved 1.2 million samples in a study by Sanzick et al.,[3] who conducted a GWAS meta-analysis for eGFR using data from CKDGen and UK Biobank and identified 634 independent signals and 23 significant genes. However, Sanzick et al's work was predominantly based on individuals of European ancestry. For East-Asian population, the largest study thus far comprised 150,266 individuals from Biobank Japan.[4]

For reviewer's reference, we highlight our new findings as follows:

(1) A meta-analysis was conducted on Japanese and Taiwanese populations. The results were validated on an independent Taiwanese cohort. Functional annotation analysis and fine mapping were used to annotate the results, which reported the discovery of *F12* and *ABCG2* as novel genes associated with kidney function in patients of Asian ancestry.

(2) A polygenic risk score (PRS) derived from the meta-analysis successfully differentiated CKD risk in two independent datasets: the CMUH-CRDR (China medical university hospital - clinical research data repository cohort) and the UK BioBank. This confirms its external validity in both East Asian and Caucasian populations. Our findings indicate that individuals with a high genetic burden for low eGFR (PRS > 2 SD) develop CKD significantly sooner than the those with a low genetic burden (PRS < -2 SD), with an onset difference of 8 years in the CMUH dataset and 3 years in the UKB. The PRS is a valuable tool for ultra-early prevention of CKD, especially in the East Asian population.

For the manuscript title, we agreed with the reviewer and changed the manuscript title as follows to remove the "ESKD-prevalent" description and provide the exact number of participants.

Revised contents: (Revised part is highlighted with underline.)

1) **TITLE, page 1, line 1:**

Original: "Discovery and prioritization of genetic determinants of kidney function in approximately 310,000 individuals from ESKD-prevalent Taiwan and Japan"

Revision: "Discovery and prioritization of genetic determinants of kidney function in 297,355 individuals from Taiwan and Japan"

2. ***The choice to seemingly report raw numbers of SNPs that were significant or replicated (e.g. lines 129 and 141) is confusing. There is no real depth of pursuit of the majority of lists of genes reported.***

Response:

We thank the reviewer for the comments. We have added lead loci to clarify the independent significant SNPs. To respond to the lack of depth of pursuit of the reported variants, we had conducted fine mapping, colocalization with eQTL, pathway analysis, and functional analysis to validate the reproducibility of identified genes in the original manuscript submitted. We included these findings in Results section (fine mapping, page 10, line 194-214; eQTL colocalization, page 11, line 214-244) and provided a detailed discussion in the Discussion section, along with a summary table of consistently highlighted candidate genes across various validation analyses in **Supplementary Table 15**. For instance, *F12* gene was validated across seven analyses: replication, correlation with BUN, clinical validation in CMUH-CRDR CKD and ESKD cohort, fine mapping for exonic SNPs and regulatory SNPs, and *cis*-eQTL colocalization.

Revised contents: (Revised part is highlighted with underline.)

1) **RESULTS, page 7, line 117-118:**

   Original: "Out of the 3,899 replicated SNPs, 387 were genome-wide significant in the TWB-based replication data set."

   Revision: "Out of the 3,899 replicated SNPs, 387 from 10 independent loci were genome-wide significant in the TWB-based replication data set."

2) **RESULTS, page 7, line 129-130:**

   Original: "The effect sizes of 2,064 replicated SNPs associated with both eGFR, and BUN were strongly …"

Revision: "The effect sizes of 2,064 replicated SNPs from 31 independent loci associated with both eGFR, and BUN were strongly…"

3) **RESULTS, page 8, line 141:**

Original: "…; these included 160 SNPs that are likely related to kidney function, …"

Revision: "…; these included 160 SNPs from 14 independent loci that are likely related to kidney function, …"

4) **RESULTS, page 8, line 149-150:**

Original: "…, including 69 likely kidney-relevant SNPs close to *KNG1*, …"

Revision: "…, including 69 likely kidney-relevant SNPs from 7 loci close to *KNG1*, …"

3. The Bonferroni correction on lines 153-154 is incorrect and it is concerning that the authors did not pick up on the fact that the corrected p-value they report was almost genome-wide significant for far fewer tests.

Response: We thank the reviewer for the opportunity to make the correction. To account for 119 genetic correlation results from 131 multiple tests, we applied the Bonferroni correction to the *p*-values. The correct significance threshold was set at 4.2×10^{-4}, which is equivalent to 0.05/119. We also revised the strength of genetic correlations of other phenotypes to eGFR and BUN GWAS meta-analysis based on the correlation coefficient, rather than the *p*-value. All statistical results are detailed in Supplementary Table 8 and are also available on the figshare website under

the specified identifier (doi:10.6084/m9.figshare.24356587). We have revised the manuscript accordingly as follows:

Revised contents: (Revised part is highlighted with underline.)

1) **RESULTS, page 9, line 153-161:**

Revision: "…, we explored the genetic correlation (r_g) of eGFR and BUN with 71 quantitative and 48 binary phenotypes from the TWB discovery dataset (Supplementary Table 8). We identified 13 genetic correlations for eGFR and 2 genetic correlations for BUN that were statistically significant ($p < 4.2 \times 10^{-4} = 0.05/119$; Supplementary Table 8). The strongest genetic correlations observed between eGFR and biomarkers, except for serum creatinine (S-Cre), were those between eGFR and BUN ($r_g = -0.30$; $p = 1.13 \times 10^{-8}$), urinary albumin ($r_g = 0.25$; $p = 5.92 \times 10^{-7}$), uric acid ($r_g = -0.24$; $p = 7.15 \times 10^{-10}$), and muscle mass ($r_g = -0.14$; $p = 1.67 \times 10^{-7}$; Supplementary Fig. 3a). For BUN, the strongest genetic correlation observed was that with serum creatinine ($r_g = 0.28$; $p = 4.06 \times 10^{-10}$; Supplementary Fig. 3b).

4. Performing targeted coloc with only significant SNPs rather than a full coloc between the GWAS and the eQTL data is debatable.

Response: Thank you for this comment. In this study, we used all SNPs from eGFR GWAS and *cis*-eQTLs within ± 100 kb regions of the lead SNPs for the colocalization analyses of each of the 111 eGFR lead SNPs, as described in Methods (page 28, line 583-592). Our methodology mirrored that of Wuttke et al.,[5] conducting colocalization analysis for each index SNP of eGFR with *cis*-eQTLs from the NEPTUNE study (focused on human glomerular and tubule-interstitial kidney tissues) and GTEx Project v6p (encompassing 44 human tissues). This would facilitate

future comparisons across similar GWAS studies on kidney function. Wuttke et al also used tissue-gene pairs with eQTL data within ± 100 kb of GWAS index SNP and tested in the regions of *cis*-eQTL windows. The manuscript has been revised to enhance clarity as follows:

Revised contents: (Revised part is highlighted with underline.)

1) **RESULTS, page 11, line 216-218:**

   Original: "For each eGFR-associated SNP, we conducted colocalization analyses by using *cis*-eQTL data from gene expression across 51 tissues, these data were sourced from GTEx v8 and the Human Kidney eQTL Atlas….

   Revision: "We conducted colocalization analyses for 111 eGFR-associated lead SNPs by using *cis*-eQTL data and eGFR GWAS within ± 100 kb regions of the lead SNPs. The *cis*-eQTLs across 51 tissues were sourced from GTEx v8 and the Human Kidney eQTL Atlas…"

5. ***And it also doesn't have face validity that the "The most intriguing discovery is that the majority of SNPs associated with decreased kidney function in East Asian populations differ from those associated with decreased kidney function in Caucasian populations." There are many other reasons for this observation which have profoundly different implications.***

Response: We acknowledge the reviewer's concerns and have removed the statement in response. In our study, we found that 8 loci were significant exclusively in the East-Asian population, 424 were significant only among individuals of European ancestry, and 89 were shared between the two groups. For instance, rs4715491 in the *FAM83B* gene and rs4148155 in the *ABCG2* gene were associated with eGFR exclusively in Asians and not in Europeans. However, we agree with the reviewer that genetic variations across ancestries can arise from

multiple factors, such as population-specific variations (e.g., founder effects) and shifts in allele frequency due to genetic drift and local selection[6], and the interaction between genetic backgrounds and environment exposures (G x E interactions).[7] We have carefully reviewed the manuscript to ensure that statements regarding this perspective are not overstated. We also included an example in the revised manuscript to prompt readers to interpret our findings with caution, as follows:

Revised contents:

1) **DISSCUSION, page 17, line 348-362:** We replaced the original paragraph with the discussions about possibility of the genetic variations as follows:

   Revision: "Our study's novel findings revealed that 8 loci were significant exclusively in the East-Asian population, 424 were significant only among individuals of European ancestry, and 89 were shared between the two groups. For instance, rs4715491 in the *FAM83B* gene and rs4148155 in the *ABCG2* gene were associated with eGFR exclusively in Asians and not in Europeans. These new findings should be approached cautiously, as genetic variations across ancestries may result from several factors. These include population-specific variations like founder effects, changes in allele frequency due to genetic drift and local selection[55], and interactions between genetic backgrounds and environmental exposures (G x E interactions).[56] The *ABCG2* gene illustrates this point well. Its rs4148155 allele frequency is 0.319 in East Asians and 0.104 in Europeans as per GnomAD[57], indicating that a larger European sample than previously used is needed for statistical significance due to its lower frequency. The effect of ABCG2 may also be amplified in Taiwan's purine-rich diet, facilitating gout and hyperuricemia phenotypes – both conditions known to predispose individuals to CKD. However, this concern underscores the importance of utilizing local genetic data to create feasible PRS for disease prevention and ultra-early screening in specific populations. Future research … "

6. ***Finally, the data availability statement is not in line with most work published in this era. At the very least, GWAS summary statistics and coloc statistics should be openly shared. Individual level data with disease age of onset and eGFR should also be able to be shared to a federated public resource without compromising patient privacy.***

Response: Thank you. We have made our summary-level data and all statistics data publicly available. The GWAS summary data that support the findings of this study are available in figshare with the identifier (doi:10.6084/m9.figshare.24356587) and statistics data are available in the supplementary document.

Revised contents: (Revised part is highlighted with underline.)

1) **DATA AVAILABILITY, page 30, line 637-639:**

   Revision: "The GWAS summary data that support the findings of this study are available in figshare with the identifier (doi:10.6084/m9.figshare.24356587) and statistics data are available in the supplementary document. The individual-level data that support the findings of this study are available on request from the corresponding author (C.-C. Kuo). The data are not publicly available due to them containing information that could compromise research participant privacy."

Reviewer: 2

Comments to the Author

1. ***Data availability: If possible, it would be beneficial to provide access to summary-level meta-analysis results of eGFR and BUN, this should not threaten privacy of participants.***

Response: Thank you. We have made our summary-level data and all statistics data publicly available. The GWAS summary data that support the findings of this study are available in figshare with the identifier (doi:10.6084/m9.figshare.24356587) and statistics data are available in the supplementary document.

Revised contents: (Revised part is highlighted with underline.)

1) **DATA AVAILABILITY, page 30, line 637-639:**

Revision: "The GWAS summary data that support the findings of this study are available in figshare with the identifier (doi:10.6084/m9.figshare.24356587) and statistics data are available in the supplementary document. The individual-level data that support the findings of this study are available on request from the corresponding author (C.-C. Kuo). The data are not publicly available due to them containing information that could compromise research participant privacy."

2. ***Kidney cell types: It could potentially be interesting to use LDsc-seg with kidney cell type data sets. You already applied this approach to investigate regulatory elements. This might help prioritize genes specifically expressed in certain kidney cell types.***

Response: We appreciate the reviewer's comments. We had incorporated the LDSC-seg method from Finucane et al. into our analysis, which includes cell type–specific gene expression.[8] The LDSC-seg method, also known as stratified LD score regression, is a tool for identifying relevant cell types based on GWAS summary statistics. We have explored the tissue- or cell type-

specificity of eGFR GWAS results using the gene expression dataset in Finucane's study. The datasets encompass kidney tissues from the GTEx project (RNA-seq data of 53 human cell types)[9] and from the Franke lab[10, 11] (Array data of 152 human/mouse cell types). The stratified LD score regression underscored the kidney tissue (A05.810.453.kidney) and kidney cortex as significant kidney tissues in the analysis. The results have been incorporated into the revised Results section and Supplementary Table 11. We have also included an additional paragraph on LDSC-seq in the Method section.

Revised contents: (Revised part is highlighted with underline.)

1) **RESULTS, page 10, line 182-193:**

   Revision: "We employed stratified LD score regression to estimate the contributions of cell type–specific functional genomic elements and tissue-specific gene expression to heritability by using GWAS summary statistics related to eGFR and BUN. Cell type–specific functional genomic elements were sourced from the Roadmap Epigenomics database[18], while tissue-specific gene expression data were derived from the GTEx[19] and Franke lab databases[20,21]. In the eGFR GWAS, fetal kidney was the enriched tissue for functional genomic elements, and the most significant tissues for tissue-specific gene expression were observed in kidney tissues (A05.810.453.kidney) and kidney cortex, followed by fetal kidney (Supplementary Table 11). In the BUN GWAS, the enriched tissue for functional genomic elements was fetal kidney and the kidney cortex was the enriched cell type for tissue-specific gene expression (Supplementary Table 11). The results suggest that kidney-specific epigenomic elements and gene expression contribute to the heritability of both eGFR and BUN."

2) **Methods, page 27, line 555-566:** We added an extra paragraph about LDSC-seq to the Method section as follows:

   Revision: "**Cell type-specific enrichment by stratified LD score regression**

We applied stratified LD score regression to identify relevant tissues and cell types pertinent to eGFR and BUN GWAS meta-analysis results, partitioning heritability from GWAS summary statistics to sets of cell type-specific regulatory elements and gene expression.[81,82] We obtained the pre-calculated EAS LD score and partitioned LD score datasets from Finucane et al's study[82], available at their LDSC resources website (https://alkesgroup.broadinstitute.org/LDSCORE/). The regulatory annotation datasets for cell type-specific analysis included 220 cell types, featuring four histone marks H3K4me1, H3K4me3, H3K9ac and H3K27ac.[83,84,85] The gene expression datasets for cell type-specific analysis were sourced from GTEx[19], comprising RNA-seq data of 53 human cell types, and the Franke lab[20,21], including array data of 152 human/mouse cell types. We set the significance threshold at an FDR (False Discovery Rate) of less than 5%.

3. Figure 3: Luckily, there are no green dots, please adjust caption.

Response: We thank the reviewer's careful review. We have revised the caption to acknowledge the absence of green dots in the figure. Also, the legend for Figure 3 has been updated to exclude any mention of green dots.

Revised contents:

1) **Figure 3, page 44:**

   Revision: "Yes" indicated by blue dots: the effect is consistent, and p is <0.05; ~~"No" indicated by green dots: the effect direction is opposite and p < 0.05;~~ "Inconclusive" indicated by grey dots: $p \geq 0.05$.

4. Table 1: Is the last pathway significant (p-adj > 0.05) or should it be removed?

Response: We appreciate the reviewer's question. We concur with the reviewer that the last pathway in Table 1 is not significant (p-adj > 0.05) and have accordingly removed it from Table 1.

5. ***Methods, line 456 - isn't this b37 => b38?***

Response: Thank you. We used GRCh37 as the reference genome for all our genotype data and analyses. However, the TWBv2 array raw data were annotated with GRCh38. To ensure consistency, we converted the TWBv2 data from GRCh38 to GRCh37 using LiftOver, a standard tool for converting genomic coordinates from one assembly to another.[12] We did not make any changes to address this response.

Reference

1. Bello AK OI, Levin A, Ye F, Saad S, Zaidi D, Houston G, Damster S, Arruebo S, Abu-Alfa A, Ashuntantang G, Caskey FJ, Cho Y, Coppo R, Davids R, Davison S, Gaipov A, Htay H, Jindal K, Lalji R, Madero M, Osman MA, Parekh R, See E, Shah DS, Sozio S, Suzuki Y, Tesar V, Tonelli M, Wainstein M, Wong M, Yeung E, Johnson DW. ISN–Global Kidney Health Atlas: A report by the International Society of Nephrology: An Assessment of Global Kidney Health Care Status focussing on Capacity, Availability, Accessibility, Affordability and Outcomes of Kidney Disease. *International Society of Nephrology*, (2023).

2. Sollis E*, et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res* **51**, D977-D985 (2023).

3. Stanzick KJ*, et al.* Discovery and prioritization of variants and genes for kidney function in >1.2 million individuals. *Nat Commun* **12**, 4350 (2021).

4. Sakaue S*, et al.* A cross-population atlas of genetic associations for 220 human phenotypes. *Nat Genet* **53**, 1415-1424 (2021).

5. Wuttke M*, et al.* A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat Genet* **51**, 957-972 (2019).

6. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26-31 (2019).

7. Tremblay J, Hamet P. Environmental and genetic contributions to diabetes. *Metabolism* **100S**, 153952 (2019).

8. Finucane HK*, et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat Genet* **50**, 621-629 (2018).

9. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-660 (2015).

10. Pers TH*, et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat Commun* **6**, 5890 (2015).

11. Fehrmann RS*, et al.* Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat Genet* **47**, 115-125 (2015).

12. Hinrichs AS*, et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res* **34**, D590-598 (2006).

Reviewer: 3

Comments to the Author

1. ***Many technical words in the manuscripts are non-standard in the field of human genetic epidemiology. The English writing is too problematic.***

Response:

We appreciate the reviewer's comments regarding the manuscript's terminology. We revised the manuscript thoroughly and confirm that all the terminology aligns with recent genetics publications. Additionally, based on the reviewer's concerns the manuscript has undergone English editing by a professional editing service to ensure clear and concise communication. We have uploaded the certificate provided by the English editing service.

2. ***ABSTRACT -"ancestral diversity."***

Response:

We appreciate the reviewer highlighting the importance of accurate terminology. The use of race, ethnicity, and ancestry in genomics research has been a topic of discussion (Kamariza *et al*., 2021). In this context, when referring to Europeans, Africans, and East Asians, the terms "ancestry" is used rather than "ethnicity". We have followed the approach outlined in the Wuttke *et al*.'s paper, using "ancestral diversity" instead of "ethnic diversity", as listed below (Wuttke *et al*., 2019). Therefore, we retained "ancestral diversity" and did not make any changes.

Reference:

- Kamariza M, Crawford L, Jones D, Finucane H. Misuse of the term 'trans-ethnic' in genomics research. *Nat Genet* **53**, 1520-1521 (2021).
- Wuttke M, *et al.* A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat Genet* **51**, 957-972 (2019).

3. ABSTRACT - the first sentence in the abstract is very strange.

Response:

Thank you for your feedback. We agree that the first sentence could be clearer. Based on the reviewer comments, we revised the sentence as follows: In genome-wide association studies (GWASs), limited representation of ancestral diversity in kidney function limits the generalizability of related findings. The East-Asian population is relatively under-represented in previous GWASs on kidney function, despite having the highest global burden of end-stage kidney disease (ESKD).

Revised contents: (Revised part is highlighted with underline.)

1) ABSTRACT, page 3, line 39-42:

Original: "Limited ancestral diversity in genome-wide association studies (GWASs) on kidney function confines their generalizability to regions with a high incidence of end-stage kidney disease (ESKD)."

Revision: "In genome-wide association studies (GWASs), limited representation of ancestral diversity in kidney function limits the generalizability of related findings. The East-Asian population is relatively under-represented in previous GWASs on kidney function, despite having the highest global burden of end-stage kidney disease (ESKD)."

4. ABSTRACT - "238 independent signals": statistical independent?

Response:

We appreciate the reviewer's attention to detail. "Statistically independent" more precisely describes the identified signals in GWAS. While our manuscript uses "independent signals" to align with the terminology in Wuttke *et al*. and Waranaba *et al*.'s studies, we acknowledge that

"statistically independent " is the more rigorous term. We revised the sentence to “238 statistically independent signals”.

Reference:

- Wuttke M, *et al.* A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat Genet* **51**, 957-972 (2019).
- Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).

Revised contents: (Revised part is highlighted with underline.)

1) **ABSTRACT, page 3, line 45-46:**

   Original: “…, which identified 238 independent signals in 97 loci.”

   Revision: “…, which identified 238 <u>statistically</u> independent signals in 97 loci.”

5. ABSTRACT - line 43-45: grammar problem

Response:

Thank you for highlighting the grammar problem. We have corrected the grammatical error in lines 43-45.

Revised contents: (Revised part is highlighted with underline.)

1) **ABSTRACT, page 3, line 46-48:**

   Original: “Functional enrichment analyses revealed that variants associated with *F12* expression and *ABCG2* missense mutation link inflammation, coagulation, and urate metabolism to a high risk of decreased kidney function.”

Revision: "Functional enrichment analysis revealed that variants associated with *F12* expression and a missense mutation in *ABCG2* correlated inflammation, coagulation, and urate metabolism exhibited elevated risk of reduced kidney function."

6. ABSTRACT - we do not say "substantially differentiate" (line 46) to express classification ability of a predictor

Response:

Thank you for comment. We have replaced "substantially differentiated" with "significantly stratified".

Revised contents: (Revised part is highlighted with underline.)

1) ABSTRACT, page 3, line 48-51:

Original: "Polygenic risk scores (PRSs) for chronic kidney disease (CKD) substantially differentiated the CKD risk in independent cohorts from Taiwan (n = 25,345) and the United Kingdom (n = 260,245)."

Revision: "In independent cohorts from Taiwan (n = 25,345) and the United Kingdom (n = 260,245), polygenic risk scores (PRSs) for chronic kidney disease (CKD) significantly stratified the risk of CKD ($p < 0.0001$)."

7. INTRODUCTION - line 55, "definitive cure": not standard

Response:

Thank you for comment. We revised the sentence to: "No cure has yet been developed for chronic kidney disease (CKD), …".

Revised contents: (Revised part is highlighted with underline.)

1) **INTRODUCTION, page 4, line 58:**

Original: "A definitive cure for chronic kidney disease (CKD) is lacking."

Revision: "No cure has yet been developed for chronic kidney disease (CKD), …"

8. INTRODUCTION - line 61, "the identification of genetic variants that cause CKD is crucial for accelerating drug development": from this study, you can at most find susceptible loci rather than finding the causal loci for CKD.

Response:

We agree that GWAS typically identifies susceptibility loci, not definitive causal variants, for complex diseases like CKD. We revised the sentence to address this concern: "identifying the genetic variants associated with CKD is essential for accelerating drug development efforts.".

Revised contents: (Revised part is highlighted with underline.)

1) **INTRODUCTION, page 4, line 64-65:**

Original: "the identification of genetic variants that cause CKD is crucial for accelerating drug development"

Revision: "identifying the genetic variants associated with CKD is essential for accelerating drug development efforts."

9. INTRODUCTION - Line 63, "More than 250 genetic loci associated with biomarkers of CKD": biomarker is a very broad term that in genetic association studies it is usually reserved to indicate variants rather than phenotype

Response:

We acknowledge that "biomarker" is a broad term. However, in the context of genomics research, it encompasses not only genetic variants but also phenotypic markers. Wuttke *et al.*

employed eGFR, a phenotypic measure, as a biomarker for quantifying kidney function in their trans-ancestry GWAS meta-analysis of CKD. Similarly, Karjalainen *et al*. utilized circulating metabolic biomarkers to represent metabolic traits in their GWAS study. Nevertheless, to address the reviewer's concern, we revised the sentence to: "More than 250 genetic loci associated with kidney functional markers of CKD, including the estimated glomerular filtration rate (eGFR) and the urine albumin-to-creatinine ratio, ….".

Revised contents: (Revised part is highlighted with underline.)

2) INTRODUCTION, page 4, line 66-67:

Original: "More than 250 genetic loci associated with biomarkers of CKD, including estimated glomerular filtration rate (eGFR) and the urine albumin-to-creatinine ratio (uACR), ..."

Revision: "More than 250 genetic loci associated with <u>kidney functional markers</u> of CKD, including the estimated glomerular filtration rate (eGFR) and the urine albumin-to-creatinine ratio, …."

Reference:

- Wuttke M, *et al.* A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat Genet* **51**, 957-972 (2019).
- Karjalainen MK, *et al.* Genome-wide characterization of circulating metabolic biomarkers. *Nature* **628**, 130-138 (2024).

10. INTRODUCTION - Line 65, "reproducible": not an appropriate word here.

Response:

We agreed with reviewer and revised sentence to address this concern: "…, are highly replicated in large-scale population-based cohorts, …".

Revised contents: (Revised part is highlighted with underline.)

1) **INTRODUCTION, page 4, line 67-68:**

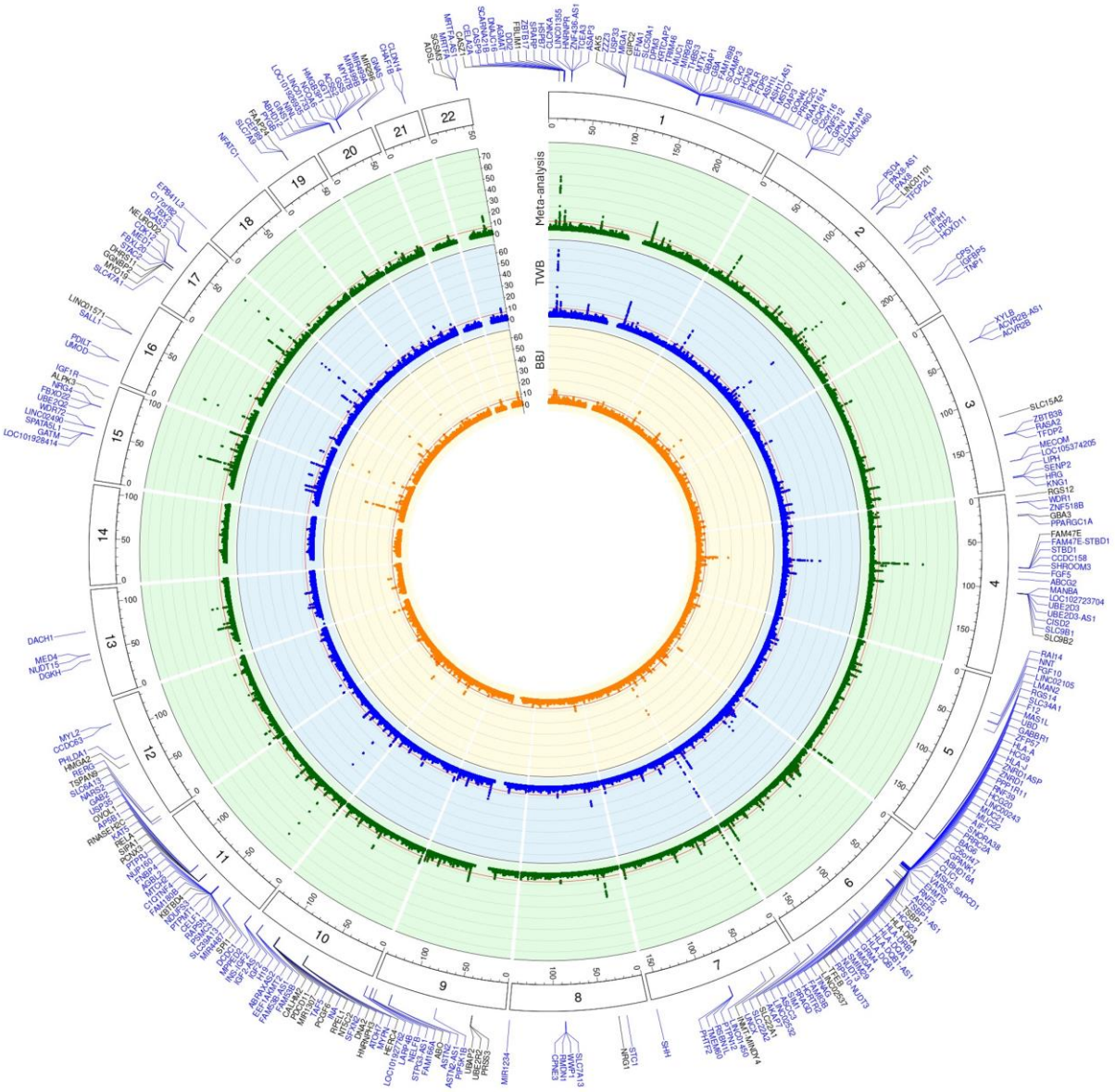
   Original: "…, are highly reproducible in large-scale population-based cohorts, …"

   Revision: "…, are highly replicated in large-scale population-based cohorts, …"

11. RESULTS - It is not very often to see that a meta-analysis discovers over 5000 genome-wide significant markers. As a standard way of presentation, the authors should output a table in the main result (not supplement) including: the summary statistic (effect size, p-value, maybe MAF) of the top variants in each dataset, and the meta-analysis summary statistic. The readers will need to check the consistency of the effect size, and interpret the increased power that generates a smaller p-value.

Response:

We appreciate the reviewer's suggestion for increased transparency in reporting our GWAS results. To address this, we included effect sizes and *p*-values for significant SNPs from each dataset and the meta-analysis in Supplementary Table 2. In addition, we generated a new circular Manhattan plot for readers to access detailed information of each dataset in Figure 2 (as shown in the below). The method used to create the circular Manhattan plot is described in the GWAS method section.



Revised contents: (Revised part is highlighted with underline.)

1) Figure 2, page 44:

Original: "**Figure 2. Manhattan plot from a meta-analysis of eGFR-derived GWASs (TWB, n = 101,294; BBJ, n = 143,658).** In total, 5,790 SNPs had p values of $<5 \times 10^{-8}$ (solid red line), and 4,732 of them exhibited a consistent effect direction. The nearest genes to SNPs were labelled. The lowest *p* value was for the SNP near UNCX. Total sample size = 244,952."

Revision: **"Figure 2. <u>A circular</u> Manhattan plot from a meta-analysis of eGFR-derived GWASs (TWB, *n* = 101,294; BBJ, *n* = 143,658; total *n* = 244,952).**

The green band corresponds to $-\log_{10}(p)$ for association with eGFR in the meta-analysis by chromosomal position. The blue band corresponds to $-\log_{10}(p)$ for association with eGFR in the TWB-derived discovery data set by chromosomal position. The orange band corresponds to $-\log_{10}(p)$ for association with eGFR in the BBJ data set by chromosomal position. The solid red line indicates genome-wide significance ($p = 5 \times 10^{-8}$). Genes labeled in black indicate SNPs exclusively identified in the meta-analysis, whereas genes labeled in blue indicate SNPs identified in the meta-analysis and additionally detected in the TWB or BBJ or both. A total of 5,790 SNPs had p values of $<5 \times 10^{-8}$, of which 4,732 had a consistent effect direction. The lowest p value was observed for rs62435145 near *UNCX* on chromosome 7 ($p = 5.23 \times 10^{-67}$; Supplementary Table 2)."

12. RESULTS - The PRS includes age and gender variable. I wonder that apart from the age, how well can the PRS based on pure genetic information can differentiate the high risk and low risk population. The study also analyzed that the heritability estimated for eGFR is only 10.9%. The PRS is constructed based on the eGFR variants to predict the CKD outcome, so I would expect that the predictability of genetic risk factor is not very strong but the power of PRS is mainly contributed by age. If that is the case, the suggested PRS as a major conclusion of the study may not be very specific for CKD or eGFR.

Response:

We appreciate the reviewer's comments. Following the reviewer's suggestion, we removed age as a covariate during the PRS model development and assessed its impact on the model's classification ability. The results showed that excluding age slightly reduced the model fit (R^2) of the PRS model from 0.013 to 0.012 (Figure R1) during the PRS model development. However, the p-value threshold for SNP selection in the optimal PRS model remained unchanged (0.0485001). In other words, the PRS model without age includes the same SNPs, resulting in no

difference in the performance between the two PRS models. The results of the classification ability are presented in Figure R2.

Figure R1. Bar chart of R^2 distribution for PRS models.

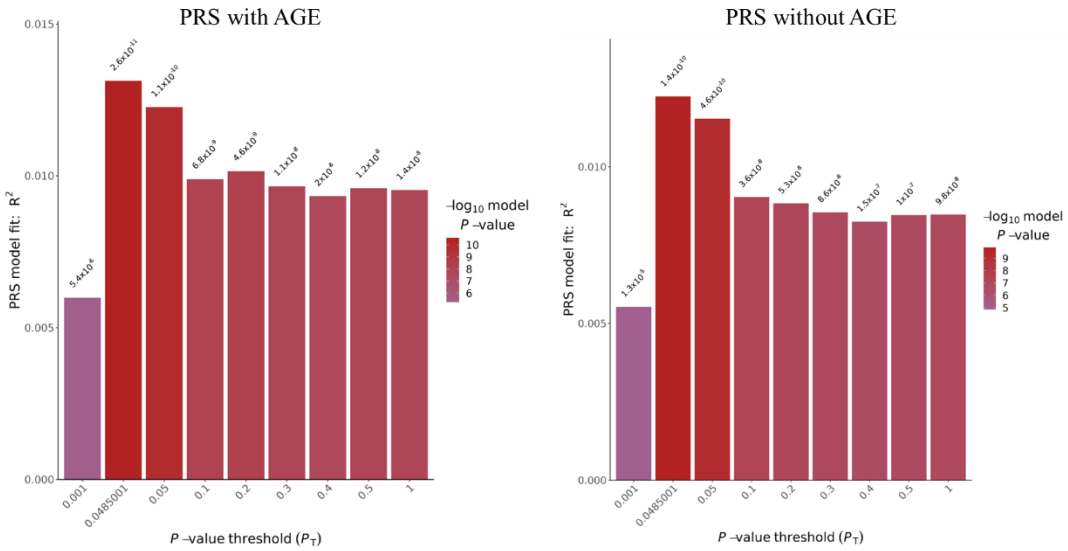
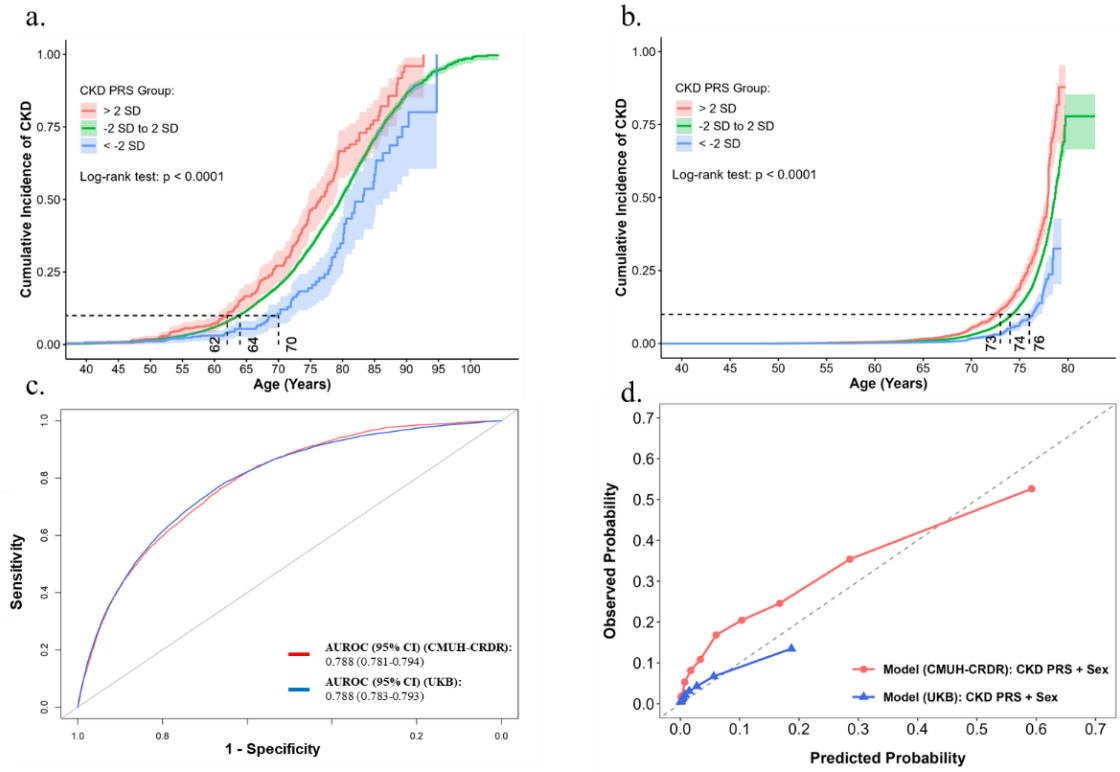


Figure R2. PRS model performance without age in the CMUH-CRDR dataset and the UKB white British dataset.

The high PRS group had a higher cumulative incidence of CKD than did the low PRS group across ages in (a) the CMUH-CRDR dataset and (b) the UKB white British dataset. The dashed line represents a cumulative incidence of CKD at 10%, which is an approximate estimate of the global prevalence of CKD. AUROC (c) and the calibration curve (d) of the PRS model are shown in both datasets.

Abbreviations*:* AUROC area under the receiver operating curve; CI, confidence interval; CKD, chronic kidney disease; CMUH-CRDR, China Medial University Hospital-Clinical Research Data Repository; PRS, polygenic risk score; SD, standard deviation; UKB, UK Biobank.



13. METHODS - BBJ: Population in this data should be directly described. This information is guessed from the mentioning of "University of Tokyo".

Response:

We appreciate the reviewer's comments. We have included information on the BBJ population in the main text of the manuscript. The BBJ comprised 143,658 available eGFR and 139,817 available BUN with a mean age (SD) of 62.9 (±13.2) years, a mean eGFR (SD) of 79.93 (±15.42) mL/min/1.73 m^2, and a mean (SD) BUN level of 15.44 (±4.77) mg/dL; 54.9% of this group were male. For further details on the eGFR and BUN populations within the BBJ cohort, please refer to Supplementary Table 1 (information adapted from the Kanai *et al*. paper). We also described TWB and UKB populations under the methods section.

Revised contents: (Revised part is highlighted with underline.)

1) METHODS, BBJ section, page 20, line 414-417:

Original: “In the current study, we used the summary data from 199,982 participants (143,658 with eGFR) for the discovery of SNPs associated with eGFR and BUN phenotypes.”

Revision: “The BBJ comprised 143,658 available eGFR and 139,817 available BUN from patient population with a mean age (SD) of 62.9 (±13.2) years, a mean eGFR (SD) of 79.93 (±15.42) mL/min/1.73 m^2, and a mean (SD) BUN level of 15.44 (±4.77) mg/dL; 54.9% of this group were male.”

2) **METHODS, TWB section, page 20, line 424-431:**

Original: “In the present study, we obtained the TWBv1 genotyping data of 27,737 individuals (27,058 with eGFR measurement) for our replication data set and the TWBv2 genotyping data from 103,332 individuals (101,294 with eGFR measurement) for our discovery data set.”

Revision: “In the present study, in addition to demographic and phenotypic data, we obtained TWBv2 genotyping data for 101,294 individuals for our discovery data set. The TWB-based discovery data set revealed a mean age (SD) of 50.4 (±10.9) years, a mean eGFR (SD) of 101.95 (±14.75) mL/min/1.73 m^2, and a mean BUN level (SD) of 13.07 (±3.90) mg/dL.; 45.6% of the individuals were male. We also obtained TWBv1 genotyping data for 27,058 individuals for our replication data set. The TWB-based replication data set revealed a mean age (SD) of 49.26 (±11.10) years, a mean eGFR (SD) of 101.83 (±15.22) mL/min/1.73 m^2, and a mean BUN level (SD) of 13.26 (4.00) mg/dL; 49.9% of the individuals were male.”

3) **METHODS, UKB section, page 21, line 441-444:**

Original: “In the current study, we used the data from 260,245 white British participants as the validation population for the predictive performance of PRS_{CKD}.”

Revision: “In this study, we used data for 260,245 White British participants as validation data to confirm the predictive performance of PRS_{CKD}. This group had a mean age (SD) of 54.98 (±18.20) years and 8,821 cases of CKD, and 45.7% of them were male.”

Reference:

- Kanai M, *et al.* Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat Genet* **50**, 390-400 (2018).

14. METHODS - Use of CMUH data is unclear. The paragraph only described the clinical variables in the database, but did not include the genetic data, yet the paragraph is titled "genetic data from CMUH". Does the author mean that the Taiwan biobank can matched to the subject in this database? In that case, you should say "match" rather than "integrate" in line 434. The ultimate source of the "genetic data from CMUH" is still unclear. Population included in this data should also be described.

Response:

The genetic data from CMUH was independently collected by CMUH and did not match that of the Taiwan Biobank. In addition to collecting clinical phenotypic data from the CMUH-CRDR, CMUH also conducted its Precision Medicine Project to gather genetic data using the TWBv2 genotyping array (see **Genotyping and imputation**). The CMUH's Precision Medicine Project consecutively enrolled outpatient participants since 2018 (Liu *et al.*, 2021). By integrating these datasets, we were able to identify 25,345 patients with eGFR measurements from the combined CMUH-CRDR (clinical data) and CMUH's Precision Medicine Project (genotyping data) cohort. This cohort had a mean age (SD) of 56.72 (±15.73) years, 4,509 cases of CKD, and 706 cases of ESKD; 44.3% of the cohort was male.

Revised contents: (Revised part is highlighted with underline.)

1) **METHODS, CMUH-CRDR section, page 22, line 450-455:**

   Original: "We integrated clinical phenotypic data from the CMUH-CRDR with TWBv2 genotyping array data for 25,345 patients with eGFR measurements, which were collected by CMUH's Precision Medicine Initiative.[63]"

Revision: "CMUH also conducted its Precision Medicine Project to gather genetic data using the TWBv2 genotyping array. The CMUH's Precision Medicine Project consecutively enrolled outpatient participants since 2018. A total of 25,345 patients with eGFR measurements were identified from the combined CMUH-CRDR (clinical data) and CMUH's Precision Medicine Project (genotyping data) cohort. This cohort had a mean age (SD) of 56.72 (±15.73) years, 4,509 cases of CKD, and 706 cases of ESKD; 44.3% of the cohort was male."

Reference:

- Liu TY, *et al.* Comparison of multiple imputation algorithms and verification using whole-genome sequencing in the CMUH genetic biobank. *Biomedicine (Taipei)* **11**, 57-65 (2021).

15. METHODS - Phenotype identification: the summary statistic of the phenotype variables, eGFR and BUN, should be provided.

Response:

We have added summary statistics for the phenotype variable, eGFR and BUN of each dataset in Method section and Supplementary Table 1. Please see the Method section of BBJ (page 20, line 414-417), TWB (page 21, line 425-431), UKB (page 21, line 441-444), CMUH (page 22, line 450-455).

16. METHODS - Line 498, any reference to support the cutoff of LD = 0.6 to define "independent SNPs". It may still not accurate to claim them as "independent" unless they are statistically independent (by definition).

Response:

We adopted the definition of independent SNPs from Watanabe *et al*.'s 2017 paper. As defined by the paper, independent significant SNPs with a genome-wide significant *p*-value ($< 5\times10^{-8}$) and independent from each other at $r^2 < 0.6$ are identified.

Reference:

- Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).

17. METHODS - Line 499: this sentence is logically not correct.

Response:

We have revised the sentence as pointed out by the reviewer: "Genomic risk loci were defined as regions with a window size of ±250 kb centered on independently significant SNPs.".

Revised contents: (Revised part is highlighted with underline.)

2) *METHODS*, page 24, line 516-517:

Original: "Genomic risk loci were defined as nonoverlapping regions if they had a window size of ±250 kb around independent significant SNPs."

Revision: "Genomic risk loci were defined as regions <u>with</u> a window size of ±250 kb <u>centered on</u> independently significant SNPs."

18. METHODS - Any reference to support the "merging" of "loci"? not a standard treatment.

Response:

The merging of loci is adopted from Watanabe *et al*'s 2017 paper. Here is the description from the paper: "Additionally, if LD blocks of independent significant SNPs are closely located to each other (< 250 kb based on the most right and left SNPs from each LD block), they are merged into one genomic locus. Each genomic locus can thus contain multiple independent significant

SNPs and lead SNPs." Merging loci has been widely accepted to guarantee the robustness of the fine mapping and signal alignment across populations (Gao *et al*., 2024).

Reference:

- Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).
- Gao B, Zhou X. MESuSiE enables scalable and powerful multi-ancestry fine-mapping of causal variants in genome-wide association studies. *Nat Genet* **56**, 170-179 (2024).

19. METHODS - Line 519: defining a marker as "replicated" in a second cohort by p-value < 0.05 is debatable. Better to directly report the p-values in the replication dataset.

Response:

We adopted a similar approach for SNP identification as used by Wuttke *et al*. (2019) and Stanzick *et al*. (2021). These studies employed directionally consistent and nominally significant associations (p-value < 0.05) to identify candidate SNPs in secondary or replication datasets. All statistical information of replicated SNPs were in Supplementary Table 4.

Reference:

- Wuttke M, *et al.* A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat Genet* **51**, 957-972 (2019).
- Stanzick KJ, *et al.* Discovery and prioritization of variants and genes for kidney function in >1.2 million individuals. *Nat Commun* **12**, 4350 (2021).

20. METHODS - Line 525: please explain why negative effect on BUN is considered as relevant to kidney function. Strong biological assumption is made here on SNP effect?

Response:

Urea, a key waste product of the liver during the breakdown of dietary protein and worn-out tissues plays a crucial role in nitrogen excretion. This nitrogen-containing molecule travels through the bloodstream to the kidneys, where it's filtered out and eliminated via the urine. Blood urea nitrogen (BUN) specifically measures the concentration of nitrogen within urea circulating in the blood (Finco *et al.*, 1976). BUN levels are inversely correlated with kidney function, meaning higher BUN levels often indicate reduced kidney function (Dossetor *et al*., 1966). Like the approach used by Wuttke et al., we can further evaluate the association of our identified genetic signals with BUN. This alternative biomarker of kidney function, inversely correlated with eGFR, helps discern whether these genetic signals predominantly reflect kidney function or merely influence creatinine metabolism.

Reference:

- Finco DR, Duncan JR. Evaluation of blood urea nitrogen and serum creatinine concentrations as indicators of renal dysfunction: a study of 111 cases and a review of related literature. *J Am Vet Med Assoc* **168**, 593-601 (1976).
- Dossetor JB. Creatininemia versus uremia. The relative significance of blood urea nitrogen and serum creatinine concentrations in azotemia. *Ann Intern Med* **65**, 1287-1299 (1966).
- Wuttke M, *et al.* A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nat Genet* **51**, 957-972 (2019).

Reviewer: 3

Comments to the Author

1) The English of the manuscript is still not okay. There are problems in nearly every 2 sentences. For instance,

Line 53, "An 8-year difference between high and low genetic risk groups was observed from the age of 50." – not sure what the author means.

Line 60, "Dapagliflozin is the only oral drug approved by the U.S. Food and Drug Administration to slow the progression of CKD, and this therapeutic effect was an unexpected discovery."

Response:

We appreciate the reviewer's efforts to help us improve the quality of our writing. We have carefully revised the manuscript sentence by sentence to ensure the clarity this time. Additionally, we have removed the unclear sentence, "An 8-year difference between high and low genetic risk groups was observed from the age of 50" from line 53 of the abstract. We also made slight modifications to the sentence at Line 60 to avoid emphasizing only one drug of the SGFLT2 inhibitors. The revised sentence is as follows:

"Sodium-glucose cotransporter-2 (SGLT2) inhibitors are the only oral drugs approved by the U.S. Food and Drug Administration to slow the progression of CKD. This therapeutic effect was an unexpected discovery."

Revised contents: (Revised part is highlighted with underline.)

1) **INTRODUCTION, page 4, lines 66-68:**

   Original: "Dapagliflozin is the only oral drug approved by the U.S. Food and Drug Administration to slow the progression of CKD, and this therapeutic effect was an unexpected discovery."

Revised: "Sodium-glucose cotransporter-2 (SGLT2) inhibitors are the only oral drugs approved by the U.S. Food and Drug Administration to slow the progression of CKD. This therapeutic effect was an unexpected discovery."

2) Scientifically, the manuscript lacks novelty. It is a meta-analysis study, without new data nor new method.

Response:

We strongly disagree with the reviewer's comments. Our manuscript presents significant new data, including genetic information from patients at Taiwan's China Medical University Hospital (CMUH), and introduces new findings. Another paper using genetic information from Taiwan's CMUH was recently published in Nature Communications (Sun *et al.*, 2024). Our study is currently the largest genetic study on kidney function within East-Asian populations. We invite the reviewer to carefully review the supplementary material and study flow charts, which clearly demonstrate the novel genetic research framework employed in this study and its relevance to the onset of CKD.

Reference:

Sun TH, Wang CC, Liu TY, et al. Utility of polygenic scores across diverse diseases in a hospital cohort for predictive modeling. *Nat Commun*. 2024;15(1):3168.

3) Results: I'm also concerned with the number of reported genome-wide significant markers - 5,790 is too large. In the largest meta-analysis GWAS for Alzheimer's disease with heritability of nearly 80%, sample size is >400,000, and the number of significant markers is 29.

In a similar meta-analysis study involving UK Biobank and Finnish cohort on Pneumonia with a total sample size > 500,000, the number of independent genome-wide significant biomarker is 2, and the gene-based analysis identifies 18 genes reaching the significance level. (Campos AI, Kho P, Vazquez-Prada KX, García-Marín LM, Martin NG, Cuéllar-Partida G, Rentería ME. Genetic susceptibility to pneumonia: a GWAS meta-analysis between the UK Biobank and FinnGen. Twin Research and Human Genetics. 2021 Jun;24(3):145-54.)

Since the heritability of kidney disease is 30-75%, we would expect that with a sample size is ~250,000, the number of significant markers would also be in the range of 1-50. But the author reported 5,790! As the author used standard software, I cannot tell from the writing where went wrong in their analysis. But >5000 significant markers do not seem right.

Response:

We appreciate reviewer's comments that help improve the clarity of our paper. We would like to assure the editors and this reviewer that nothing went wrong in this analysis. The number of discovered SNPs depends on the disease phenotypes, heritability strength, and the threshold set in the analysis (Tam *et al.*, 2019; Schaid *et. al.*, 2019).

First, throughout the manuscript, we will strictly apply the terminology of genome-wide significant SNPs, independent significant SNPs, lead SNPs, and genomic risk loci, as detailed in **Table R1**. We invite the editors and this reviewer to align their understanding with these definitions, which are widely accepted in the literature.

Table R1. Terminology of different types of SNPs.

Term	Definition
Genome-wide significant SNP	SNPs achieve the genome-wide statistical significance threshold of P value $< 5 \times 10^{-8}$ (Schaid *et. al.,* 2019).
Independent significant SNP	SNPs with a genome-wide significant *P*-value and independent from each other at $r^2 < 0.6$ are identified (Karadag *et. al.,* 2023; Watanabe *et. al.,* 2017).
Lead SNP	SNPs are defined by clumping independent significant SNPs with $r^2 < 0.1$ (Meddens *et. al.,* 2021; Watanabe *et. al.,* 2017).
Genomic risk loci	When LD blocks of independent significant SNPs are closely located to each other (< 250 kb based on the most right and left SNPs from each LD block), they are merged into one genomic locus. Each genomic locus can thus contain multiple independent significant SNPs and lead SNPs (Ou *et. al.,* 2023; Watanabe *et. al.,* 2017).

Second, we have also conducted the experiments according to the papers mentioned by the reviewer and provided data on how the numbers of SNPs, using different terminologies and definitions, change. These details are provided for both editors and reviewers in the following steps:

(1) We followed reviewer's guidance and confirmed that the authors in the Alzheimer paper (Jansen et al., 2019) have used the same LD cutoff for independent variant selection as we did. Jansen *et al*. also used FUMA, an online platform for fine-mapping, to identify genome-wide significant SNPs that have a *P*-value ($< 5 \times 10^{-8}$) and independent significant SNPs are independent from each other at $r^2 < 0.6$. From these independent significant SNPs, they identified lead SNPs that met the $r^2 < 0.1$ criteria, and further defined associated genomic risk loci by merging any physically overlapping lead SNPs (LD blocks < 250 kb apart). As a result, they revealed 2,357 genome-wide significant SNPs, 236 independent significant SNPs, 94 lead SNPs, and 29 genomic risk loci for Alzheimer. Following the same method in the current study, we

identified 5,790 genome-wide significant SNPs, 283 independent significant SNPs, 111 lead SNPs, and 97 genomic risk loci for eGFR (**Table R2**).

Table R2. Comparison of different types of SNPs identified in the current study with those reported in Jansen's study.

	Jansen *et al.*'s study (Alzheimer)	**Current study (eGFR)**
Genome-wide significant SNP	2357	5790
Independent significant SNP	236	283
Lead SNP	94	111
Genomic risk loci	29	97

(2) We followed the second referenced pneumonia paper (Campos *et al*., 2021) to re-analyze. In Campos *et al.*'s study, two independent genetic signals (two lead SNPs) were identified by clumping with $r^2 < 0.05$, and within 1Mb window. Following their methods, we identified 113 lead SNPs for eGFR (**Table R3**).

Table R3. Comparing the number of lead SNPs in the current study with those reported in Campos's study.

	Campos *et al.*'s study (Pneumonia)	**Current study (eGFR)**
Lead SNPs ($r^2 < 0.05$, 1MB window)	2	113

(3) To increase the stringency, we re-run the analysis using $r^2 < 0.01$ as the LD cutoff and within 1MB window. As a result, we identified **102 lead SNPs** for eGFR.

To improve the clarity in presenting results, we modified the description to emphasize independent significant variants, lead SNPs and genomic risk loci in the abstract and the main text, and removed the total number of GWS SNPs in **Figure 1**.

Third, to further assure the editors and reviewers, we provide a brief summary of papers targeting the phenotype of eGFR with sample sizes above 100,000. The number of discovered genome-

wide significant SNPs can be reasonably large; three papers even reported significant SNPs numbering above 20,000 (**Table R4**).

Table R4. Summary of public eGFR GWAS dataset.

Database	Journal Article	Population	Sample size	Number of SNPs	Number of significant SNP (P < 5e-08)
IEU OpenGWAS	Nat Comm 2015 (PMID:26831199)	Trans-ancestry (Meta-analysis)	133,814	2,198,208	1,720
		European-non diaebtes (Meta-analysis)	118,460	2,198,288	1,700
IEU OpenGWAS	Nat Genetics 2018 (PMID: 29403010)	East Asian	143,658	5,961,601	4,001
GWAS Catalog	Nat Genetics 2019 (PMID: 31152163)	Trans-ancestry (Meta-analysis)	765,348	8,221,591	26,442
		European (Meta-analysis)	567,460	8,834,748	23,064
GWAS Catalog	Nat Comm 2021 (PMID: 34272381)	Trans-ancestry (Meta-analysis)	1,201,909	13,633,840	62,095
Biobank Japan Project	medRxiv 2021	East Asian	154,633	13,531,752	8,562
GWAS Catalog	Nat Genetics 2022 (PMID: 35710981)	Trans-ancestry (Meta-analysis)	1,508,659	12,653,804	92,545
Current Study		**East Asian (Meta-analysis)**	**244,952**	**4,512,904**	**5,790**

IEU, Integrative Epidemiology Unit; GWAS, Genome-wide association studies

In summary, we are confident that our analytical results are robust and aligned with the consensus in the existing literature.

Reference:

1. Campos AI, Kho P, Vazquez-Prada KX, et al. Genetic Susceptibility to Pneumonia: A GWAS Meta-Analysis Between the UK Biobank and FinnGen. *Twin Res Hum Genet*. 2021;24(3):145-154.
2. Jansen IE, Savage JE, Watanabe K, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat Genet*. 2019;51(3):404-413.
3. Karadag N, Shadrin AA, O'Connell KS, et al. Identification of novel genomic risk loci shared between common epilepsies and psychiatric disorders. *Brain*. 2023;146(8):3392-3403.
4. Meddens SFW, de Vlaming R, Bowers P, et al. Genomic analysis of diet composition finds novel loci and associations with health and lifestyle. *Mol Psychiatry*. 2021;26(6):2056-2069.
5. Ou YN, Ge YJ, Wu BS, et al. The genetic architecture of fornix white matter microstructure and their involvement in neuropsychiatric disorders. *Transl Psychiatry*. 2023;13(1):180.
6. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet.* 2018;19(8):491-504.
7. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. *Nat Rev Genet.* 2019;20(8):467-484.
8. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun.* 2017;8(1):1826.

Revised contents: (Revised part is highlighted with underline.)

1) **ABSTRACT, page 3, lines 48-49:**

   Original: "… and identified 238 statistically independent signals in 97 genomic loci."

   Revised: "… and identified <u>111 lead SNPs</u> in 97 genomic <u>risk</u> loci."

2) **RESULTS, page 6, lines 111-114:**

   Original: "A total of 97 nonoverlapping genomic risk loci containing 238 independently significant SNPs were identified, of which 46 were unreported…"

Revised: "A total of 238 independently significant SNPs (LD $r^2 < 0.6$) were represented by 111 lead SNPs (LD $r^2 < 0.1$) located in 97 genomic risk loci, of which 26 were unreported…"

3) **RESULTS, page 6, lines 115-120:**

Original: "Genomic risk loci were defined as nonoverlapping on the basis of a window size of ±250 kb around independently significant SNPs, and these loci were merged into a single locus if the distance between them was shorter than 250 kb. The previously unreported eGFR-associated SNP with the lowest *p* value ($p = 5.1 \times 10^{-45}$) was rs754331108, which was located in the intronic region of *CASP9*. The minor alleles across eGFR-associated SNPs both increased and decreased the eGFR, …"

Revised: "Genomic risk loci were defined as nonoverlapping on the basis of a window size of ±250 kb around LD block of lead SNPs, and these loci were merged into a single locus if the distance between them was shorter than 250 kb. The previously unreported genome-wide significant SNP with the lowest *p* value ($p = 5.1 \times 10^{-45}$) was rs754331108, which was located in the intronic region of *CASP9*. The minor alleles across genome-wide significant SNPs both increased and decreased the eGFR, …"

4) **RESULTS, page 7, lines 128-130:**

Original: "We evaluated the replication of eGFR-associated SNPs in an independent replication dataset that comprised data from 27,058 individuals in the TWB."

Revised: "We evaluated the replication of eGFR-associated SNPs, defined as genome-wide significant SNPs, in an independent replication dataset that comprised data from 27,058 individuals in the TWB."

5) **RESULTS, page 7, lines 134-136:**

Original: "Of the 3,899 replicated SNPs, 387 from 10 independent loci were genome-wide significant in the TWB-based replication dataset."

Revised: "Of the 3,899 replicated eGFR-associated SNPs, 387 from 10 independent genomic risk loci were genome-wide significant in the TWB-based replication dataset."

6) **RESULTS, page 8, lines 147-148:**

Original: "The effect sizes of 2,064 replicated SNPs from 31 independent loci associated with eGFR and BUN were strongly and inversely correlated…"

Revised: "The effect sizes of 2,064 replicated eGFR-associated SNPs from 31 independent genomic risk loci associated with eGFR and BUN were strongly and inversely correlated…"

7) **RESULTS, page 8, lines 159-160:**

Original: "These SNPs included 160 from 14 independent loci likely related to kidney function, …"

Revised: "These SNPs included 160 from 14 independent genomic risk loci likely related to kidney function, …"

8) **RESULTS, page 9, line 168:**

Original: "…69 likely kidney-related SNPs from 7 loci close to the following genes: …"

Revised: "…69 likely kidney-related SNPs from 7 genomic risk loci close to the following genes: …"

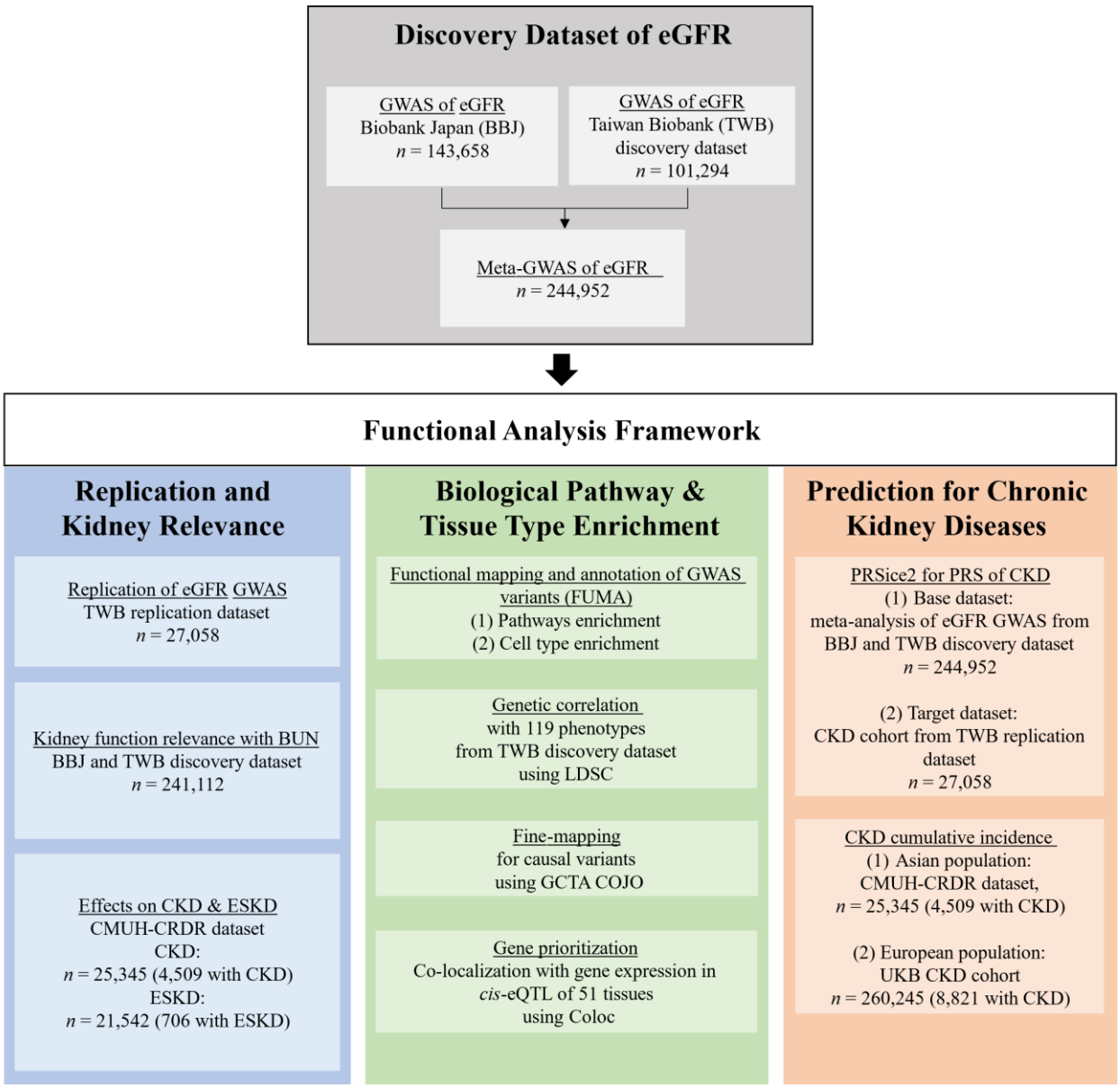
9) **RESULTS, page 11, lines 218-221:**

Original: "Subsequently, statistical fine mapping of eGFR loci was conducted through summary statistics–based conditional analyses, and 238 independent genome-wide significant SNPs mapped to 97 regions were identified."

Revised: "Subsequently, statistical fine mapping of eGFR loci was conducted through summary statistics–based conditional analyses for 238 independent significant SNPs mapped to 97 genomic risk loci."

10) **Figure 1, pages 48-49:**

We removed the total number of GWS SNPs in **Figure 1**



4) Conclusion is too strong:

Line 54, "In countries with a high incidence of ESKD, PRSCKD can be used for early kidney disease prevention." – this conclusion is too strong at this stage.

Response:

Thank you! We have revised the statement as follows:

"More research is required to replicate our findings and evaluate the clinical effectiveness of PRS_{CKD} in the early prevention of kidney disease."

Revised contents: (Revised part is highlighted with underline.)

1) **ABSTRACT, page 3, lines 58-61:**

   Original: "In countries with a high incidence of ESKD, PRS_{CKD} can be used for early kidney disease prevention."

   Revised: "More research is required to replicate our findings and evaluate the clinical effectiveness of PRS_{CKD} in the early prevention of kidney disease."

Reviewer: 4

Comments to the Author

1) In the chapter "Genetic correlations of the eGFR and BUN with other phenotypes," there is an inconsistency in the reported p-value for the correlation between BUN and S-Cre. The manuscript lists it as 4.06x10^-10, whereas Table ST8 lists it as 4.06x10^-8. Please correct this discrepancy in the manuscript or the table. Please also see Supplementary Figure 3b for this.

Response: We appreciate the reviewer for the opportunity to make the correction. The correct *p*-value for the correlation between BUN and S-Cre is 4.06×10^{-8} as shown in Table ST8. We have revised this *p*-value in both the main manuscript and Supplementary Figure 3b.

Revised contents: (Revised part is highlighted with underline.)

1) **RESULTS, page 9, lines 179-181:**

   Original: "For BUN, the strongest genetic correlation observed was that with S-Cre (r_g = 0.28, $p = 4.06 \times 10^{-10}$; Supplementary Fig. 3b)."

   Revised: "For BUN, the strongest genetic correlation observed was that with S-Cre (r_g = 0.28, $p = 4.06 \times 10^{-8}$; Supplementary Fig. 3b)."

2) **Supplementary Fig. 3b, page 4:**

Original: "**(b)** For BUN, the strongest genetic correlation observed was that with serum creatinine (r_g = 0.28; $p = 4.06 \times 10^{-10}$)."

Revised: "**(b)** For BUN, the strongest genetic correlation observed was that with serum creatinine (r_g = 0.28; $p = \underline{4.06 \times 10^{-8}}$)."

2) In the abstract, the sentence "An 8-year difference between high and low genetic risk groups was observed from the age of 50" is unclear. I recommend removing this.

Response: We acknowledge the reviewer's concerns and have removed the sentence in the abstract.

3) In the introduction, line 71, the phrase "common variations" should be revised to "common variants." Here, a "variant" refers to genetic differences such as SNPs, indels, or deletions.

Response: We appreciate the reviewer highlighting the importance of accurate terminology. We have revised the phrase "common variations" to "common variants".

Revised contents: (Revised part is highlighted with underline.)

1) **INTRODUCTION, page 4, lines 77-79:**

   Original: "…, and extensive linkage disequilibrium (LD) across common variations in these loci…"

   Revised: "…, and extensive linkage disequilibrium (LD) across <u>common variants</u> in these <u>kidney function-related</u> loci…"

4) In the discussion, last paragraph, line 396, the mention of "2,140 novel SNPs" appears without prior context. I could not find any reference to this number elsewhere in the manuscript. Please clarify where this figure originates and provide the relevant details in the Results section, or remove this from the discussion.

Response: We thank for the reviewer's comment. The "2,140 novel SNPs" refer to the genome-wide significant SNPs labeled as "No" in the Reported column in Table ST2. The unreported genome-wide significant SNPs were defined as genome-wide significant SNPs which were not identified in the eGFR GWASs from over one million individuals (Wuttke et al., 2019 and Stanzick et al., 2021). We have now added this information and relevant details to the Results section for better context and clarity. In order to maintain consistency, we revised "novel SNPs" to "previously unreported genome-wide significant SNPs".

Reference:

1. Wuttke M, Li Y, Li M, et al. et al. A catalog of genetic loci associated with kidney function from analyses of a million individuals. *Nature Genet*. 2019;51(6):957-972.
2. Stanzick KJ, Li Y, Schlosser P, et al. Discovery and prioritization of variants and genes for kidney function in >1.2 million individuals. *Nat Commun*. 2021;12(1):4350.

Revised contents: (Revised part is highlighted with underline.)

1) **DISCUSSION, page 22, lines 460-462:**

   Original: "In the present large eGFR-based GWAS of an EAS population, we identified 2,140 novel SNPs related to kidney function."

   Revised: "In <u>this</u> GWAS of <u>eGFR using large data exclusively from an East Asian</u> population, we identified 2,140 <u>previously unreported genome-wide significant</u> SNPs."

2) **RESULTS, page 6, lines 110-111:**

Original: “As shown in Fig. 2 and Supplementary Table 2, a meta-analysis of these GWASs revealed 5,790 genome-wide significant ($p < 5 \times 10^{-8}$) single-nucleotide polymorphisms (SNPs).”

Revised: “As shown in Fig. 2 and Supplementary Table 2, a meta-analysis of these GWASs revealed 5,790 genome-wide significant ($p < 5 \times 10^{-8}$) single-nucleotide polymorphisms (SNPs). In these genome-wide significant SNPs, 2,140 were unreported in the previous eGFR GWASs.”