# The Impact of Ancestry on Genome-Wide Association Studies

Steven Christopher Jones[1]*, Katie M. Cardone[2]*, Yuki Bradford[2], Sarah A. Tishkoff[2,4], Marylyn D. Ritchie[2,3,5]

*[1]Genomics and Computational Biology Graduate Group, [2]Department of Genetics, [3]Institute for Biomedical Informatics, [4]Department of Biology, [5]Department of Biostatistics, Epidemiology, and Informatics*

*University of Pennsylvania, Perelman School of Medicine, Philadelphia, PA, USA*

**Equal contributions to the manuscript*

*Email: marylyn@pennmedicine.upenn.edu*

Genome-wide association studies (GWAS) are an important tool for the study of complex disease genetics. Decisions regarding the quality control (QC) procedures employed as part of a GWAS can have important implications on the results and their biological interpretation. Many GWAS have been conducted predominantly in cohorts of European ancestry, but many initiatives aim to increase the representation of diverse ancestries in genetic studies. The question of how these data should be combined and the consequences that genetic variation across ancestry groups might have on GWAS results warrants further investigation. In this study, we focus on several commonly used methods for combining genetic data across diverse ancestry groups and the impact these decisions have on the outcome of GWAS summary statistics. We ran GWAS on two binary phenotypes using ancestry-specific, multi-ancestry mega-analysis, and meta-analysis approaches. We found that while multi-ancestry mega-analysis and meta-analysis approaches can aid in identifying signals shared across ancestries, they can diminish the signal of ancestry-specific associations and modify their effect sizes. These results demonstrate the potential impact on downstream post-GWAS analyses and follow-up studies. Decisions regarding how the genetic data are combined has the potential to mask important findings that might serve individuals of ancestries that have been historically underrepresented in genetic studies. New methods that consider ancestry-specific variants in conjunction with the shared variants need to be developed.

*Keywords:* GWAS; Ancestry; Health Disparities.

## 1. Introduction

### 1.1. *Population Structure in Genome-Wide Association Studies*

Genome-wide association studies (GWAS) are a powerful tool for discovering genetic associations with traits of interest[1]. Since its introduction in 2005, the use of GWAS has become a standard method in the field of statistical genetics, offering insight into the contribution of alleles with small effect sizes for complex traits[2]. As DNA sequencing becomes more affordable, and large healthcare systems, biobanks, and consortia continue to link electronic health record (EHR) information containing disease phenotypes to patients' genetic information, larger sample sizes for complex disease show continued promise for the application of GWAS. At the time of writing this manuscript, the GWAS catalog contained summary statistics for over 5,000 phenotypes[3].

Beyond the wide application of GWAS in the field of genetics, considerable work has been done to identify the impact of quality control (QC) procedures and best practices for GWAS[4,5]. Technical decisions such as allele frequency threshold, variant quality thresholds, data missingness, and population structure are all known to impact GWAS outcomes[5]. Despite the considerable work that has been done to offer guidance on GWAS QC and study design, many decisions are made on a case-by-case basis and the approach taken can vary based on the lab and the guidance referenced[1,4,5]. We aim to focus specifically on the impact that different strategies for combining genetic data from two genetically inferred ancestry groups have on GWAS summary statistics.

An individual's genetic ancestry can be inferred from their DNA, which contains information about the genetic signatures resulting from ancestral migrations, mutations, recombination, genetic drift, and natural selection[4,6,7]. Ancestry-specific evolutionary and demographic histories can lead to linkage disequilibrium (LD) and allele frequencies that differ across populations and result in spurious associations due to the confounding effects of ancestry in GWAS[8,9]. Some standard methods to control for population structure within a GWAS study cohort are the use of a mixed model combined with a genetic relationship matrix (GRM), principal component analysis (PCA), and the subsequent inclusion of a small number of principal components (PCs) as covariates in the GWAS model[10,11]. However, even with the inclusion of PCs, population structure may not be entirely accounted for, leading to persistent spurious associations[12]. Additional methods of inferring genetic ancestry such as K-means clustering and quadratic discriminant analysis (QDA) of PCA data or the application of tools such as ADMIXTURE can provide greater resolution for decisions regarding the inference of genetic ancestry of individuals and prove useful for QC decisions for GWAS in admixed and multi-ancestry cohorts[13,14].

As the volume of genetic data combined with rich EHR phenotype data from diverse populations continues to increase, GWAS will continue to be an important tool. Subsequently, the choice between a study focused on ancestry-specific and/or multi-ancestry GWAS approaches will have important implications on the results and their interpretations, especially when GWAS summary statistics are used for downstream analyses such as transcriptome-wide association studies (TWAS), proteome-wide association studies (PWAS), or polygenic scores (PGS)[4,15,16]. Ancestry-specific GWAS may provide insight into genetic associations within specific ancestral groups, allowing for the detection of associations that may be unique or have varying effect sizes across different populations. However, these approaches can be limited due to smaller sample sizes in underrepresented global populations. Multi-ancestry mega-analysis GWAS or meta-analysis approaches can leverage larger sample sizes and provide insight into genetic associations shared across ancestrally diverse populations[4,15,17]. However, both approaches present unique challenges and opportunities that must be carefully considered in the experimental design and interpretation of results.

### 1.2. *Inclusion of Diverse Ancestries in Genetic Studies*

Genetic studies are predominantly focused on European ancestry, with most GWAS conducted in these populations, leading to insights that are not always generalizable to non-European groups and exacerbate health disparities[3,17–22]. The lack of diversity in genetic research limits our understanding

of genetic variation in underrepresented ancestries and its relationship with complex traits[19,21]. Initiatives like the All of Us Research Program, the Human Heredity and Health in Africa (H3Africa) Initiative, the Million Veteran Program (MVP), and the NHLBI Trans-Omics for Precision Medicine program (TOPMed) aim to address this by recruiting diverse populations and creating more representative datasets for genetic research[22–25]. However, integrating these diverse datasets into GWAS is complicated by unequal sample sizes and differences in allele frequency and LD patterns between populations, which highlight the need for robust and specialized methodologies to ensure accurate and equitable interpretation of genetic associations.

Incorporating diverse ancestries in GWAS offers opportunities to discover associations absent in European-focused studies, providing valuable insight for underrepresented populations[16,26]. It can also enhance fine mapping by leveraging genomic diversity across ancestries[17]. Multi-ancestry mega-analysis and ancestry-specific GWAS with meta-analysis offer solutions but are limited by differences in study design, sample sizes, and the model specified for the meta-analysis. Decisions between fixed effect or random effect meta-analysis will have an impact on the combined results and require assumptions regarding the heterogeneity of associations between populations[4,27,28].

### 1.3. *Shared and Ancestry-Specific Associations*

Most human genetic variation can be observed within all ancestry groups and many genetic associations with disease are shared across human populations[29]. However, for a small portion of the genome, associations can vary across different ancestral populations, with distinct loci contributing to the same trait in populations with distinct genetic ancestry. This is evident in Solomon Islanders, where a mutation in the *TYRP1* gene is associated with blond hair[30]. This mutation is absent outside of Oceania, and thus cannot explain blond hair in individuals of European ancestry[30]. Similarly, variants such as the G1 and G2 variants in *APOL1* have been shown to account for a substantial degree of risk for chronic kidney disease (CKD) in individuals of African ancestry while being very rare or absent in other ancestry groups[31–33]. These examples underscore the importance of conducting ancestry-specific GWAS to uncover genetic associations that may be masked, diluted, or even missing in multi-ancestry analyses.

Many GWAS of complex traits have identified associations that are shared across ancestries in which a shared variant demonstrates a similar effect size for a trait across multiple populations[26]. For example, variants in the *FTO* gene have been consistently associated with increased body mass index across diverse populations[34]. Similarly, variants in the *TCF7L2* gene are strongly associated with increased risk of type 2 diabetes (T2D) across multiple populations[35–44].

The basis of phenotypic variation and the influence of genetic ancestry is complex. Some diseases exhibit ancestry-specific genetic associations, while others share common genetic associations across populations. This complexity is further compounded by the continuous nature of admixture in natural populations. Understanding the genetic factors that influence complex traits across different populations is crucial for developing personalized medicine approaches tailored to the unique genetic makeup of diverse individuals. The present study aims to contribute to this understanding by investigating the genetic associations with chronic kidney disease (CKD) and type

2 diabetes mellitus (T2D) across European (EUR) and African (AFR) ancestries, utilizing both ancestry-specific and multi-ancestry GWAS approaches to comprehensively assess the impact of genetic variation on these traits (**Figure 1**).

## 2. Methods

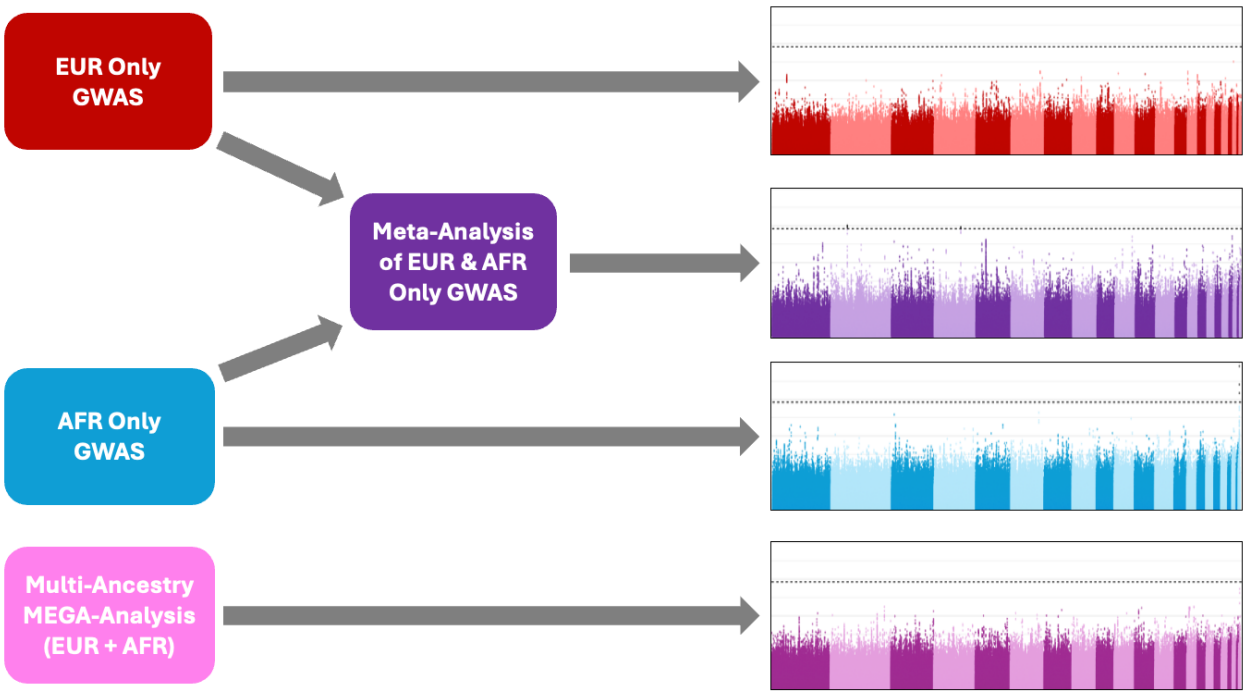


**Figure 1: Study Overview:** For each binary phenotype, four GWAS were run: EUR-specific, AFR-specific, EUR and AFR combined (multi-ancestry mega-analysis), and meta-analysis of EUR- and AFR-specific GWAS.

### 2.1. *Data and Study Participants*

The Penn Medicine BioBank (PMBB) is an electronic health record (EHR)-linked research program at the University of Pennsylvania, Perelman School of Medicine[45]. PMBB participants provided consent for research, including blood sample collection, generation of genetic data, and EHR access[45]. Individuals with imputed genotype, demographic, and EHR data were included in this study. PMBB v2.0 imputed data and v2.3 phenotype data were utilized[45].

### 2.2. *PMBB Centralized Genotyping, Imputation, & Quality Control*

DNA was extracted from blood samples, which were genotyped by the Regeneron Genomics Center with an Illumina Global Sequencing Array v2.0 (GSAv2) containing 654,027 fixed markers[45]. Variant and sample-level quality control was conducted prior to genotype imputation using PLINK v1.9[45,46]. Variants with genotype call rates < 95%, individuals with discordance between reported sex and genetic sex, and individuals with sample call rates < 90% were dropped[45]. Subsequently, autosomes were imputed using TOPMed version R2 genome build 38 reference panel[25,45,47]. After

imputation, PLINK v2.0 was used for additional variant and sample-level quality control[45,46]. Variants with genotype call rates < 99%, minor allele frequency (MAF) < 1%, Hardy-Weinberg Equilibrium (HWE) exact test p-value < 1e-8 or imputation $R^2$ scores < 0.3 were excluded[45]. Palindromic SNPs, insertions and deletions, and multiallelic variants were also excluded. In addition, individuals with sample call rates < 99% were dropped[45].

## 2.3. *Principal Component Analysis, Genetically Inferred Ancestry, and Ancestry-Specific Quality Control*

### 2.3.1 *Quality Control Prior to Principal Components Analysis*

Prior to PCA, quality control was conducted in all eligible samples using PLINK v1.9 and v2.0[46]. Individuals with sample call rates < 95% were dropped[46]. In addition, variants with genotype call rates < 95%, imputation $R^2$ scores < 0.80, MAF < 5%, or HWE exact test p-value < 1e-10 were excluded[46]. Subsequently, only variants in the intersection between the PMBB and 1,000 Genomes genetic datasets were included[6].

### 2.3.2 *Principal Component Analysis and Genetically Inferred Ancestry*

Principal component analysis (PCA) was conducted with eigensoft smartPCA on the LD pruned autosomal data[48]. PCs in PMBB were projected onto 1,000 Genomes[6,48]. Using the top two PCs, genetically inferred ancestry was computed using QDA with 1,000 Genomes super-populations as a reference[6,14]. Individuals that had >80% probability of similarity to clusters representing the 1,000 Genomes super-population of EUR of AFR were retained for inclusion in GWAS.

### 2.3.3 *Analysis-Specific Quality Control and Principal Components Analysis*

After computing genetically inferred ancestry, analysis-specific quality control was completed in EUR, AFR and MEGA (union of EUR and AFR) cohorts with PLINK v1.9 and v2.0[46]. Individuals with sample call rates < 95% and variants with genotype call rates < 95%, MAF < 95%, or imputation $R^2$ score < 0.3 were excluded. Only biallelic and non-palindromic SNPs were retained. PCA was conducted within each cohort independently following QC using eigensoft smartPCA[48]. Principal components from the cohort-specific PCA were used as covariates in the GWAS.

## 2.4. *Genome Wide Association Study*

GWAS were conducted using SAIGE[11]. We conducted GWAS utilizing three stratification methods: GWAS stratified to EUR individuals only (EUR-specific), GWAS stratified to AFR individuals only (AFR-specific), and GWAS with both EUR and AFR individuals (MEGA). We tested associations with two phenotypes: CKD and T2D. To phenotype individuals, ICD-9 and ICD-10 codes were mapped to PhecodeX if they had at least two separate instances of an ICD code[49]. The Phecodes used were as follows: CKD = GU_582.2, T2D = EM_202.2[49]. Eligible controls had zero instances of an ICD code used in case definition. To mitigate the effects of sample size, we randomly down sampled while matching case control ratio to ensure the same number of cases and controls across

EUR and AFR individuals for each phenotype. The multi-ancestry mega-analysis GWAS contained a balanced number of EUR and AFR individuals, and the same total sample size as ancestry-specific GWAS. Age at data release, sex assigned at birth, and PC1-7 were used as covariates. We selected the top seven PCs because this explained 79-98% of variance between individuals in the three cohorts (**Supplementary Figure 1, Supplementary Figure 2, Supplementary Figure 3**).

### 2.5. *Meta-Analysis*

Summary statistics from the AFR and EUR ancestry-specific GWAS analyses were meta-analyzed using METASOFT[27,28]. To compare the impact of model specification on the outcome, the meta-analyses were conducted using a fixed-effect (FE), random-effect (RE), modified random-effect (RE2_INITIAL), and modified random-effect with adjustment for mean effect and heterozygosity (RE2_CORRECTED)[27,28]. Meta-analyses were conducted on the intersection of variants included in the AFR and EUR-specific GWAS. All summary statistics from independent GWAS were adjusted using genomic control following the instructions in the METASOFT publication[27,28]. To ensure consistent sample sizes between analyses, the EUR and AFR groups were randomly down sampled prior to GWAS while maintaining balanced case control ratio such that the meta-analyses contained the same total sample size as the other GWAS. GWAS and meta-analysis results were visualized using qqman and SynthesisView[50,51]. Variants that had a p-value < 5e-8 were considered significant.

### 2.6. *Analysis of Effect Size Variability*

To assess changes in effect size for variants included in all analyses, we identified whether a variant's effect size changed direction in at least one analysis. We compared effect sizes in the following analyses: all analyses, ancestry-specific compared to multi-ancestry approaches, MEGA analysis compared to meta-analysis approaches, and fixed effect meta-analysis compared to random effect meta-analysis. We identified the percentage of variants that changed direction of effect in each comparison group, both genome-wide and among the variants with the most significant associations, which were visualized in SynthesisView plots[51].

## 3. Results

The PMBB had 43,589 individuals with genetic data that passed initial QC and were analyzed using QDA to infer genetic ancestry. Using our approach, we identified 10,631 individuals that clustered with the AFR super population and 17,495 individuals that clustered with the EUR super population from the 1,000 Genomes reference panel. **Figure 2** shows the individuals from PMBB in the PCA projection of the 1,000 Genomes. Following analysis-specific QC of these individuals, there were 10,631 individuals and 6,792,866 variants in the AFR analyses, 17,495 individuals and 4,910,840 variants in the EUR analyses, and 28,126 individuals and 5,652,287 variants in the MEGA analyses. Of these variants, 4,184,455 were shared between AFR and EUR cohorts and could be included in

meta-analyses and 3,334,796 were only found in a single ancestry after QC. **Table 1** shows the final sample sizes.

The AFR-specific GWAS of CKD replicated a known signal in the *APOL1* gene (rs73885319) on chromosome 22 (p-value = 7.92e-11) (**Figure 3, Figure 4**)[31]. This signal was not detected in the EUR-specific analysis as the MAF of this variant was 0.00869% and therefore did not pass QC. This signal was detected in the MEGA analysis with a p-value of 1.43e-7, which is below the genome-wide significance threshold. Due to the monomorphic nature of this allele in the EUR population, the variant was not included in any of the meta-analyses. The meta-analyses identified additional associations in the *ANXA5* gene on chromosome 4 and downstream of *LOC124900539* on chromosome 2.

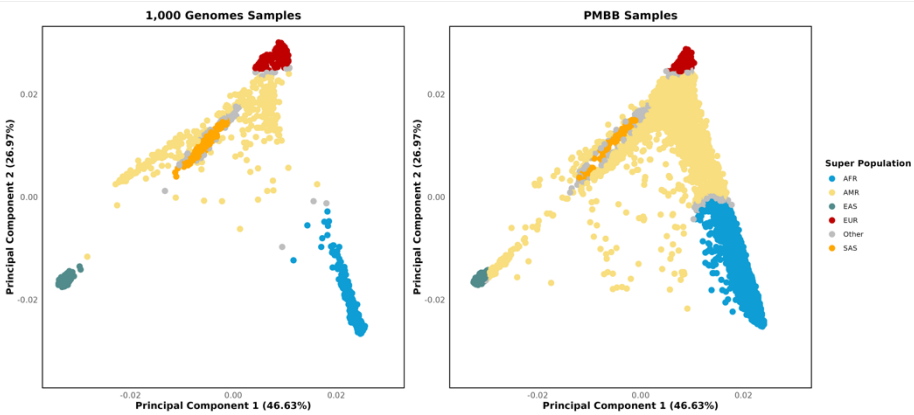


**Figure 2: PCA of PMBB samples** (right) **projected onto the 1,000 Genomes reference panel** (left). Colors indicate clustering with 1,000 Genomes super-population (AFR, AMR, EAS, EUR, SAS).

| Phenotype | Case | Control | Total Sample Size |
|---|---|---|---|
| T2D | 3,184 | 6,448 | 9,632 |
| CKD | 2,659 | 7,543 | 10,202 |

**Table 1: Final Sample Sizes for both Ancestry groups.**

The T2D GWAS replicated four known signals in the *TCF7L2* gene on chromosome 10 (rs35011184, rs7901695, rs7903146, rs34872471), and one upstream of the *CRYBA2* gene/downstream of the *MIR375* gene on chromosome 2 (rs113414093) (**Figure 3, Figure 5**)[35,36,38–44,44,52]. rs7903146 reached genome-wide significance in the AFR-specific GWAS (p-value = 6.59e-10) and the EUR-specific GWAS (p-value = 5.23e-9). This signal was detected in the MEGA and meta-analyses but was below genome-wide significance. rs34872471 was genome-wide significant in the EUR-specific GWAS (p-value = 5.00e-9) but not in the other analyses. rs35011184 and rs7901695 were detected in all GWAS iterations but were not genome-wide significant, with the EUR-specific GWAS having the lowest p-values (rs35011184 p-value = 4.05e-8, rs7901695 p-value = 1.19e-6). rs113414093 was only detected in the EUR-specific GWAS and was not genome-wide significant (p-value = 9.97e-7). This variant was not present in the other analyses as the MAF was 0.909% in the AFR-specific cohort and 3.70% in the MEGA cohort. The meta-analysis identified additional associations in the *PTPRG* gene on chromosome 3, and upstream of *LOC105374348*/downstream of *FAM53A* on chromosome 4.

In the GWAS of CKD, majority of the variants with the most significant p-values changed direction of effect in at least one analysis (**Table 2**). There was variability in the T2D analyses, but the trend was not as extreme (**Table 2**).

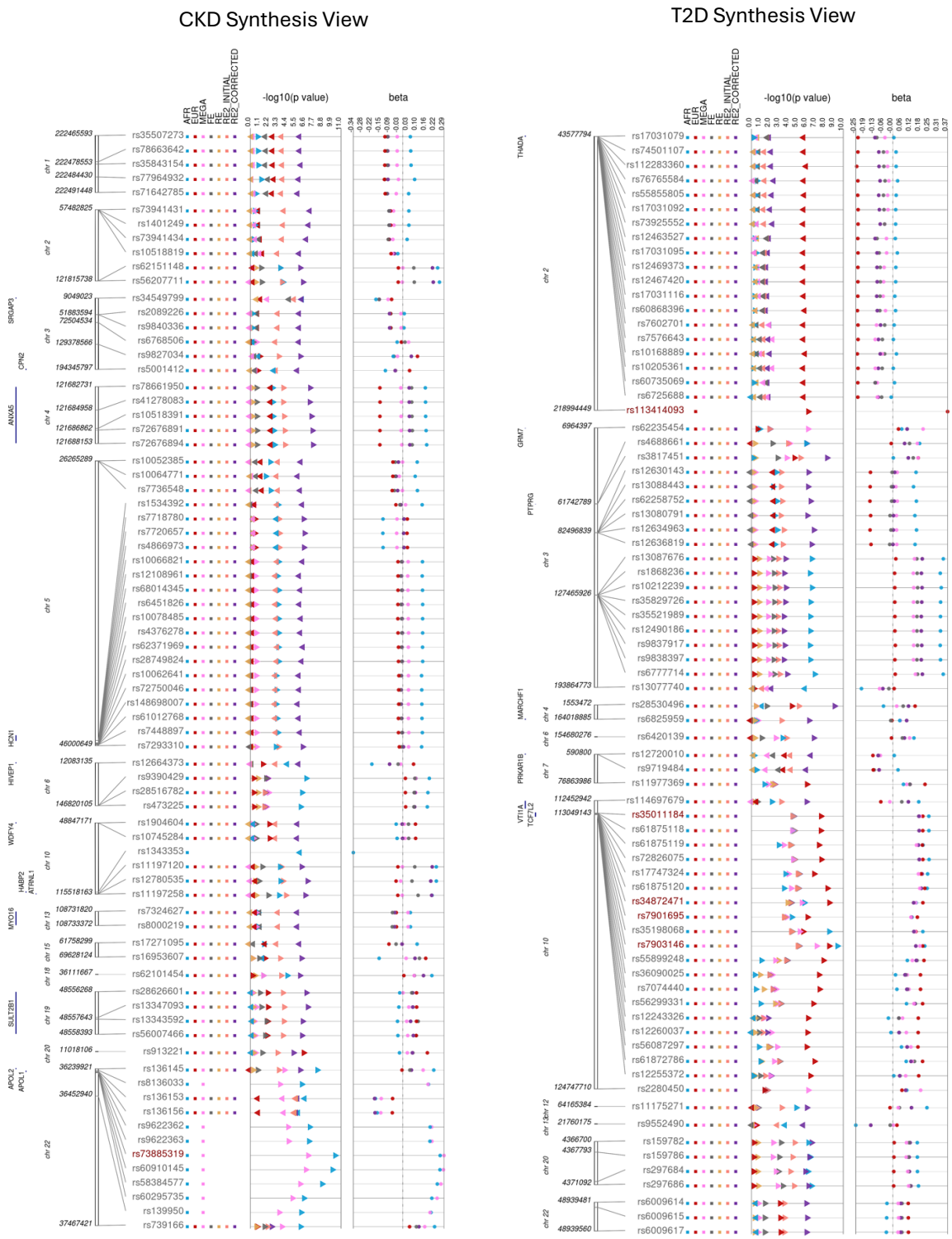


**Figure 3: Significance Levels and Effect Sizes for Chronic Kidney Disease (left) and Type 2 Diabetes (right).** The most significant variants for each phenotype are displayed. Variants highlighted in red are known signals. Variants are annotated with gene names (left axis).

| Phenotype | All Analyses | AFR vs. Multi-Ancestry Analyses | EUR vs. Multi-Ancestry Analyses | MEGA vs. Meta Analyses | Fixed Effect vs. Random Effect Meta Analyses |
|---|---|---|---|---|---|
| **Percentage of CKD Variants** | 86.36% | 80.30% | 56.06% | 50% | 7.58% |
| **Percentage of T2D Variants** | 54.05% | 47.30% | 18.92% | 12.16% | 9.46% |

**Table 2: Proportion of Top Variants that Changed Direction of Effect.** 75 variants were included in the T2D comparison, and 66 variants were included in the CKD comparison.


**Figure 4: Chronic Kidney Disease Stacked Manhattan Plot.** Top plot is AFR-specific GWAS, followed by MEGA GWAS, EUR-specific GWAS, and meta-analysis using modified random effect framework (RE2_corrected).

Additionally, direction of effect flipped less when comparing multi-ancestry methods (**Table 2**). When investigating variants genome-wide, there is a decrease in variability in CKD, but an increase in variability in T2D (**Supplementary Table 1**). Additionally, 84-98% of the most significant variants' effect sizes in multi-ancestry analyses had a value within the range of ancestry-specific effect sizes (**Supplementary Table 2**). This trend was less extreme in variants genome-wide, as

nearly 50% of effect sizes in multi-ancestry analyses were within the range of ancestry-specific effect sizes.

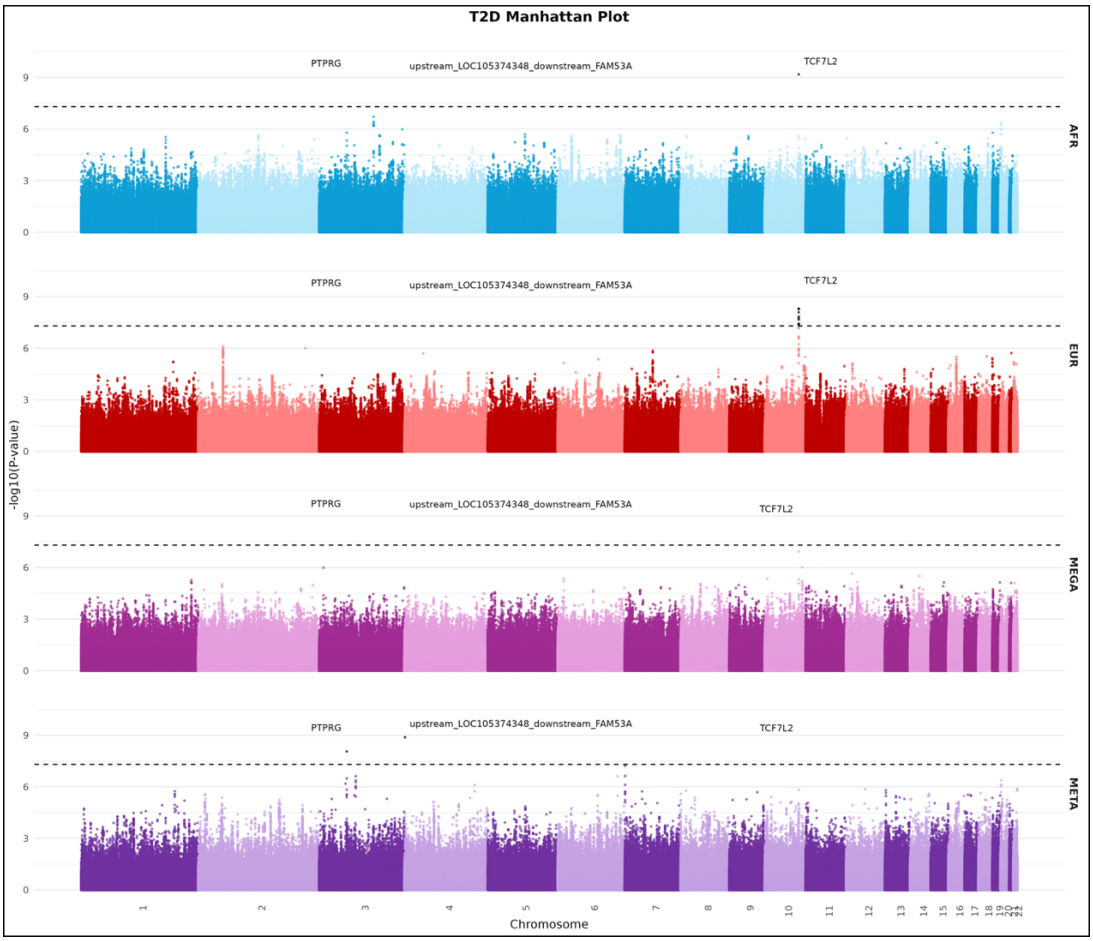


**Figure 5: Type 2 Diabetes Stacked Manhattan Plot.** Top plot is AFR-specific GWAS, followed by MEGA GWAS, EUR-specific GWAS, and ancestry-balanced meta-analysis using modified random effect framework (RE2_corrected).

## 4. Discussion

Our aim was to assess how different approaches of combining genetic data from individuals of diverse ancestries change the outcome of a GWAS. To test this, we conducted GWAS of CKD and T2D in individuals of African and European ancestry in the PMBB. We compared the differences in GWAS results through changes to the p-value and effect sizes for ancestry-specific analyses (AFR or EUR only), multi-ancestry mega-analysis (MEGA), and meta-analysis using fixed-effect (FE), random-effect (RE), and modified random-effect (RE2_INITIAL and RE2_CORRECTED). We hypothesized that while most genetic associations are shared across human populations, we would observe specific genetic associations that were statistically significant in only one ancestry and that the different multi-ancestry approaches would have inconsistent results for these variants. The results support our hypothesis as shown in **Figures 3-5** and **Table 2**.

In the GWAS of CKD, variants within the *APOL1* gene were found to be significantly associated with CKD in the AFR-specific GWAS[31]. In the mega-analysis GWAS, these variants dropped below genome-wide significance, providing evidence that multi-ancestry mega-analysis can diminish ancestry-specific signals. We also note that the use of a meta-analysis tool such as METASOFT will exclude the association observed in the AFR-specific GWAS due to this variant not passing QC in the EUR cohort. Additional variants in the *ANXA5* gene and downstream of *LOC124900539* were significantly associated in the meta-analysis (RE2_CORRECTED) but may be spurious due to genomic inflation in this approach (**Supplementary Figure 4**).

In the AFR and EUR-specific GWAS of T2D, a well-known variant (rs7903146) within the *TCF7L2* gene was significantly associated with T2D[35,43,44,52], while it dropped below genome-wide significance in all multi-ancestry analyses. The GWAS of T2D illustrates how the composition of a multi-ancestry approach can diminish the significance of ancestry-specific signals. However, we acknowledge the limitation that smaller number of cases per ancestry might have had in the multi-ancestry approaches. Additional variants in the *PTPRG* gene and upstream of *LOC105374348*/ downstream of *FAM53A* were significantly associated in the meta-analysis (RE2_CORRECTED) but may be spurious due to genomic inflation in this approach (**Supplementary Figure 5**).

Across both phenotypes, effect sizes flipped direction on many occasions, especially among variants with the lowest p-values (**Table 2, Supplementary Table 1**). This occurred more often when comparing ancestry-specific approaches to multi-ancestry approaches, rather than within multi-ancestry approaches, suggesting that observed ancestry-specific effect sizes can be altered when using multi-ancestry GWAS approaches. Additionally, effect size values in multi-ancestry results were commonly within the range of ancestry-specific effect size value for variants with the lowest p-values (**Supplementary Table 2**).

Meta-analyses can be performed using different approaches, with fixed-effect (FE) and random-effect (RE) models being most common. Fixed-effect meta-analysis assumes a homogenous effect size between studies, meaning any variation in the observed effects is attributed solely to sampling error[27]. In contrast, random-effect meta-analysis assumed that the effect size varies between studies due to differences in population or study designs, allowing for more flexibility in capturing heterogeneity across datasets[27]. We employed the RE2 method developed by Han and Eskin (2011) because it improves statistical power by relaxing the conservative assumptions of the traditional random-effect model, enabling better detection of associations in the presence of heterogeneity[27].

Our study had several limitations. Our sample sizes were limited due to down sampling to match case and control numbers across ancestry groups, so many variants did not reach genome-wide significance. This is of particular importance when considering changes to the signal in the *TCF7L2* gene in T2D between approaches. Although a higher sample size would be ideal, down sampling was a crucial step to isolate the impact of ancestry on GWAS approaches rather than sample size and statistical power. Additionally, down sampled groups were not matched by age, sex or other clinical characteristics. In addition, the modified meta-analysis in the RE2_CORRECTED analyses produced slightly inflated results which often had the most significant associations and identified several signals for CKD and T2D that had not been reported in ClinVar or the GWAS

catalog[3,53]. Due to the low sample sizes in our study compared to previously reported GWAS of T2D and CKD that had not detected these associations, it is plausible these associations may be spurious. Our meta-analyses also only included variants that intersected between the ancestry-specific GWAS, which led to the exclusion of several important ancestry-specific signals in the meta-analysis results. This can be overcome through the inclusion of more cohorts in a meta-analysis but highlights an important limitation of the meta-analysis approach under our framework for directly comparing two studies. Additionally, our method to assess variability in effect sizes was unable to fully quantify observed variability. The pattern of sample overlap between the GWAS approaches in our study violated assumptions of independence or matched dependence between studies. Quantification of this variability using a well calibrated statistical methodology is a logical next step to investigate the differences observed between approaches.

In a typical GWAS, multi-ancestry mega-analysis, or meta-analysis approaches benefit from increased sample size. Our study, however, maintained consistent sample size across approaches to isolate ancestry's impact. We found that multi-ancestry methods can diminish ancestry-specific signals, which can significantly impact downstream analyses like TWAS, PWAS, or PGS. This raises questions about the optimal approach for generating summary statistics, as results differ in meaningful ways based on initial GWAS method. Notably, many variants show striking changes in effect direction, both among those with significant p-values and genome-wide. These effect size flips are crucial, as they influence downstream analyses and biological/clinical interpretations. While many variants show consistent results across approaches, a notable subset are impacted by the choice of analysis method. As we see with variants in *APOL1*, some of these variants showing variable results or which could not be fully assessed in all approaches are essential for understanding differences in disease risk between populations. Thus, new methods that consider the ancestry-specific variants in conjunction with the multi-ancestry shared variants need to be developed.

## 5. Acknowledgements

We acknowledge the Penn Medicine BioBank (PMBB) for providing data and thank the patient-participants of Penn Medicine who consented to participate in this research program. We would also like to thank the Penn Medicine BioBank team and Regeneron Genetics Center for providing genetic variant data for analysis. The PMBB is approved under IRB protocol# 813913 and supported by Perelman School of Medicine at University of Pennsylvania, a gift from the Smilow family, and the National Center for Advancing Translational Sciences of the National Institutes of Health under CTSA award number UL1TR001878. Additional funding support was provided to MDR by AI077505, EY023557, AG066833, and HL169458. Additional funding was provided to SAT by ADA 1-19-VSN-02, and NIH grants 1R35GM134957, R01AR076241, and 1X01HL139409-01.

## 6. Supplementary Material

All supplemental data can be found at:
**https://ritchielab.org/publications/supplementary-data/psb-2025/jonescardone**

**References**


1. Uffelmann, E. *et al.* Genome-wide association studies. *Nat. Rev. Methods Primer* **1**, 1–21 (2021).
2. Loos, R. J. F. 15 years of genome-wide association studies and no signs of slowing down. *Nat. Commun.* **11**, 5900 (2020).
3. Sollis, E. *et al.* The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* **51**, D977–D985 (2023).
4. Peterson, R. E. *et al.* Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. *Cell* **179**, 589–603 (2019).
5. Truong, V. Q. *et al.* Quality Control Procedures for Genome-Wide Association Studies. *Curr. Protoc.* **2**, e603 (2022).
6. The 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
7. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
8. Li, C. C. Population subdivision with respect to multiple alleles. *Ann. Hum. Genet.* **33**, 23–29 (1969).
9. Lander, E. S. & Schork, N. J. Genetic dissection of complex traits. *Science* **265**, 2037–2048 (1994).
10. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
11. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat. Genet.* **50**, 1335–1341 (2018).

12. Hellwege, J. *et al.* Population Stratification in Genetic Association Studies. *Curr. Protoc. Hum. Genet.* **95**, 1.22.1-1.22.23 (2017).
13. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
14. Qin, X., Lock, T. R. & Kallenbach, R. L. DA: Population structure inference using discriminant analysis. *Methods Ecol. Evol.* **13**, 485–499 (2022).
15. Rosenberg, N. A. *et al.* Genome-wide association studies in diverse populations. *Nat. Rev. Genet.* **11**, 356–366 (2010).
16. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
17. Li, Y. R. & Keating, B. J. Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations. *Genome Med.* **6**, 91 (2014).
18. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 26–31 (2019).
19. Ju, D., Hui, D., Hammond, D. A., Wonkam, A. & Tishkoff, S. A. Importance of Including Non-European Populations in Large Human Genetic Studies to Enhance Precision Medicine. *Annu. Rev. Biomed. Data Sci.* **5**, 321–339 (2022).
20. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10**, 3328 (2019).
21. Fatumo, S. *et al.* Diversity in Genomic Studies: A Roadmap to Address the Imbalance. *Nat. Med.* **28**, 243–250 (2022).
22. Bick, A. G. *et al.* Genomic data in the All of Us Research Program. *Nature* **627**, 340–346 (2024).

23. The H3Africa Consortium *et al.* Enabling the genomic revolution in Africa. *Science* **344**, 1346–1348 (2014).

24. Gaziano, J. M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).

25. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).

26. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).

27. Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am. J. Hum. Genet.* **88**, 586–598 (2011).

28. Han, B. & Eskin, E. Interpreting Meta-Analyses of Genome-Wide Association Studies. *PLOS Genet.* **8**, e1002555 (2012).

29. Rosenberg, N. A. *et al.* Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).

30. Kenny, E. E. *et al.* Melanesians blond hair is caused by an amino acid change in TYRP1. *Science* **336**, 554 (2012).

31. Genovese, G. *et al.* Association of Trypanolytic ApoL1 Variants with Kidney Disease in African-Americans. *Science* **329**, 841–845 (2010).

32. Parsa, A. *et al.* APOL1 risk variants, race, and progression of chronic kidney disease. *N. Engl. J. Med.* **369**, 2183–2196 (2013).

33. Pollak, M. R. & Friedman, D. J. APOL1 and APOL1-Associated Kidney Disease: A Common Disease, an Unusual Disease Gene − Proceedings of the Henry Shavelle Professorship. *Glomerular Dis.* **3**, 75–87 (2023).

34. Fawcett, K. A. & Barroso, I. The genetics of obesity: FTO leads the way. *Trends Genet.* **26**, 266–274 (2010).

35. Grant, S. F. A. *et al.* Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat. Genet.* **38**, 320–323 (2006).

36. Haddad, S. A. *et al.* A novel TCF7L2 type 2 diabetes SNP identified from fine mapping in African American women. *PloS One* **12**, e0172577 (2017).

37. Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885 (2007).

38. Uribe-Salazar, J. M., Palmer, J. R., Haddad, S. A., Rosenberg, L. & Ruiz-Narváez, E. A. Admixture mapping and fine-mapping of type 2 diabetes susceptibility loci in African American women. *J. Hum. Genet.* **63**, 1109–1117 (2018).

39. Chang, Y.-C. *et al.* Association study of the genetic polymorphisms of the transcription factor 7-like 2 (TCF7L2) gene and type 2 diabetes in the Chinese population. *Diabetes* **56**, 2631–2637 (2007).

40. Ng, M. C. Y. *et al.* Replication and identification of novel variants at TCF7L2 associated with type 2 diabetes in Hong Kong Chinese. *J. Clin. Endocrinol. Metab.* **92**, 3733–3737 (2007).

41. Lehman, D. M. *et al.* Haplotypes of transcription factor 7-like 2 (TCF7L2) gene and its upstream region are associated with type 2 diabetes and age of onset in Mexican Americans. *Diabetes* **56**, 389–393 (2007).

42. Spracklen, C. N. *et al.* Identification of type 2 diabetes loci in 433,540 East Asian individuals. *Nature* **582**, 240–245 (2020).

43. Zeggini, E. *et al.* Replication of Genome-Wide Association Signals in UK Samples Reveals Risk Loci for Type 2 Diabetes. *Science* **316**, 1336–1341 (2007).

44. Vujkovic, M. *et al.* Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat. Genet.* **52**, 680–691 (2020).
45. Verma, A. *et al.* The Penn Medicine BioBank: Towards a Genomics-Enabled Learning Healthcare System to Accelerate Precision Medicine in a Diverse Population. *J. Pers. Med.* **12**, 1974 (2022).
46. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
47. Hinrichs, A. S. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).
48. Patterson, N., Price, A. L. & Reich, D. Population Structure and Eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
49. Shuey, M. M. *et al.* Next-generation phenotyping: introducing phecodeX for enhanced discovery research in medical phenomics. *Bioinformatics* **39**, btad655 (2023).
50. Turner, S. D. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *J. Open Source Softw.* **3**, 731 (2018).
51. Pendergrass, S. A., Dudek, S. M., Crawford, D. C. & Ritchie, M. D. Synthesis-View: visualization and interpretation of SNP association results for multi-cohort, multi-phenotype data and meta-analysis. *BioData Min.* **3**, 10 (2010).
52. Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).
53. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980-985 (2014).