**Supplemental Material**

# Systematic heritability and heritability enrichment analysis for diabetes complications in ACCORD and UK Biobank Studies

[*]Corresponding author: Jin J. Zhou
Department of Medicine
David Geffen School of Medicine
University of California, Los Angeles
Email: jinjinzhou@ucla.edu

## Table of Contents

## UKB phenotype definition

**Myocardial infarction (MI)**. Cases of MI were identified by the International Classification of Disease, Ninth and Tenth Revision (ICD-10) code families I21, I22, and I23 (Acute, subsequent, and complications of MI). The primary source was UKB's fields 131298, 131300, and 131302 (Date I21/I22/I23 first reported, respectively). These fields gather information from hospital admissions, death records, primary care, and self-reported outcomes from surveys taken at UK Biobank assessment centers at initiation into the study and map them to three-digit ICD-10 categories. To obtain the most up-to-date information, we also gathered these ICD-10 code families directly from hospital admission and death records. We also included cases of MI identified through UKB's algorithmically defined outcome (field 42000). Controls were required to have no evidence of certain cardiovascular diseases.

**Unstable angina**. Cases of unstable angina were identified by the ICD-10 code I20.0, extracted from hospital admissions and death records.

**Ischemic stroke (Stroke infarct)**. Cases of ischemic stroke were identified in a manner similar to MI, using a combination of UKB's first occurrence field 131366 (Date I63 first reported (cerebral infarction), the algorithmically defined outcome for ischemic stroke (field 42008), and the ICD-10 code I63 in hospital admission or death records. Controls were required to have no evidence of cerebrovascular disease (ICD10 codes I6*, G45*, G46*).

**Stroke (Stroke any)**. Stroke was taken to be the first occurrence of either ischemic or hemorrhagic stroke, or of unspecified stroke via UKB fields 42006 (algorithmically defined stroke), 131368 (unspecified stroke), or ICD10 code I64. Controls were required to have no evidence of cerebrovascular disease (ICD10 codes I6*, G45*, G46*).

**Percutaneous coronary intervention (PCI)**. Cases of PCI were identified through OPCS4 codes K40, K41, K42, K43, K44, K45, K46, K483, K49, K501, K75, K76, and UKB self-report codes 1070 (coronary angioplasty) and 1095 (coronary bypass grafts). Controls were not to have self-reported any non-coronary revascularization procedures.

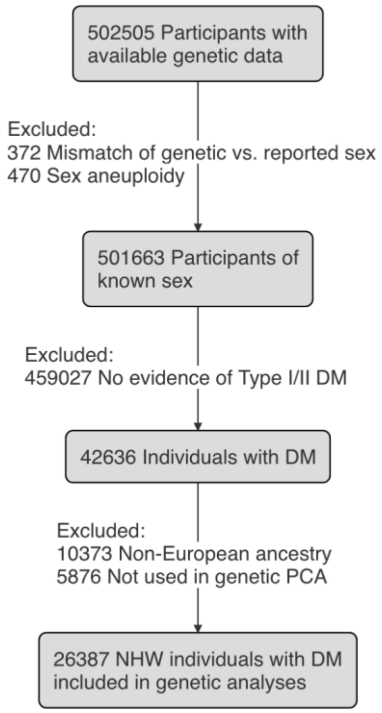
**Composite CVD (CVD)**. A composite CVD event consisted of either MI, ischemic stroke, unstable angina, or PCI. The first date of CVD was taken as the first date of any of these events. Controls were required to satisfy all the conditions for each component outcome.

**Macroalbuminuria/Microalbuminuria**. Urine Albumin:Creatinine ratio (UACR) was calculated using UKB fields 30700 (urine creatinine), 30500 (urine microalbumin), and 30505 (reason for missing urine microalbumin).  UACR above 33.9 was considered macroalbuminuria, while above 3.4 was considered microalbuminuria. In cases where urine microalbumin was below detectable levels, albuminuria status was inferred from urine creatinine where possible.

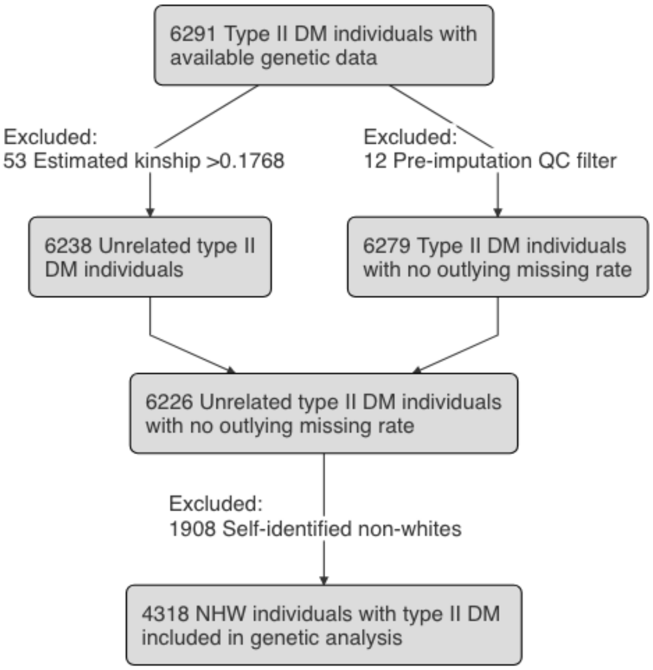
**Chronic/Diabetic kidney disease (DKD)**. DKD was identified through UKB's algorithmically defined end-stage renal disease (field 42026, previously described), ICD10 codes E1*.2 (diabetes mellitus with renal complications),  E18[0345] (chronic kidney disease stage 3-5, end-stage), N08.3 (glomerular disorders in diabetes mellitus) in hospital or death records, self-reported diabetic kidney disease, two or more consecutive eGFR (EPI creatinine) < 60 mL/min/1.73m$^2$ measured 90+ days apart from either UK Biobank Assessment Center or primary care data. The date of the first DKD was taken as the first occurrence of any of the previous codes/events. Controls were required not to have micro/macroalbuminuria or a list of exclusion codes. Controls were required to have at least five years of follow-up since their diabetes diagnosis, and cases were required to have more than five years between their date of diabetes diagnosis and first DKD.

**Diabetic eye disease (DR)**. DR was determined using the ICD10 codes E1*.3 (diabetes mellitus with ophthalmic complications), H36.0 (diabetic retinopathy), and H28.0 (Diabetic Cataract), as well as a set of primary care codes.  Since most cases were identified through primary care data, controls were required to have this data available in order to reduce misclassification. Controls were also required not to have glaucoma, cataract, or non-diabetic/unspecified retinopathy.

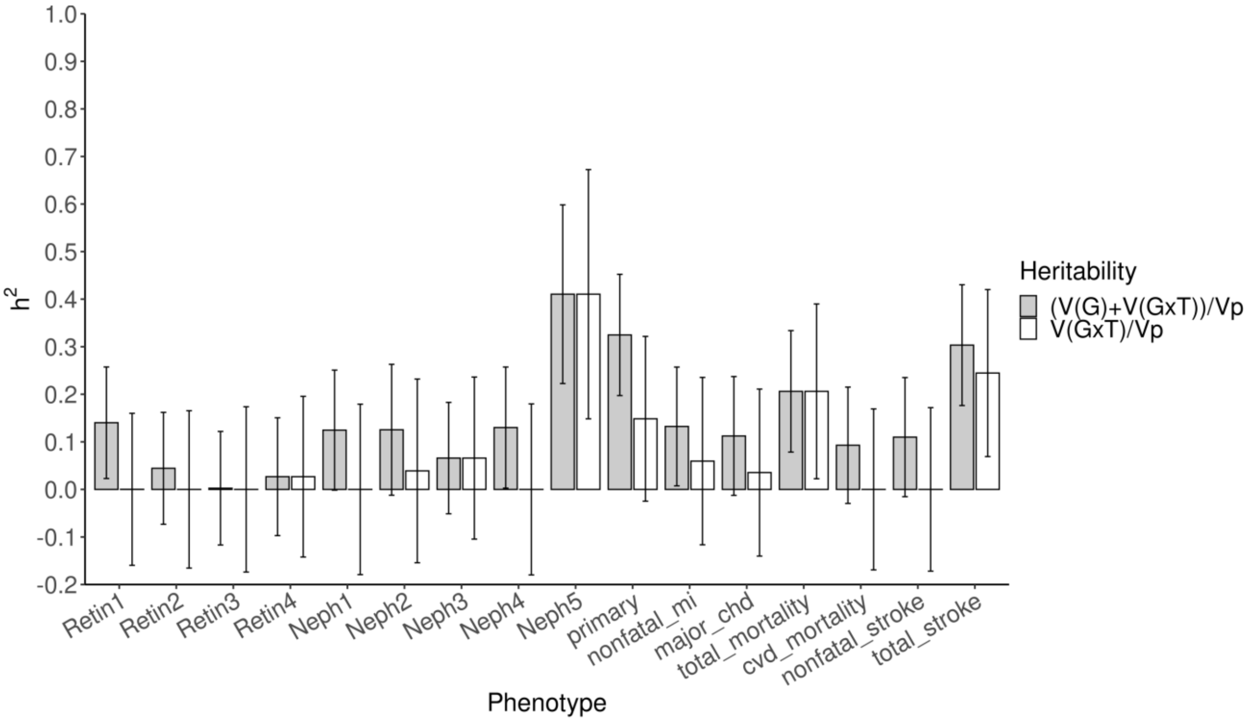
## Additional figures and tables


502505 Participants with available genetic data
Excluded:
372 Mismatch of genetic vs. reported sex
470 Sex aneuploidy
501663 Participants of known sex
Excluded:
459027 No evidence of Type I/II DM
42636 Individuals with DM
Excluded:
10373 Non-European ancestry
5876 Not used in genetic PCA
26387 NHW individuals with DM included in genetic analyses

Supplemental Figure 1. Diagram depicting a flow of participants used in the UKB analyses. DM, diabetes mellitus. NHW, non-Hispanic white.

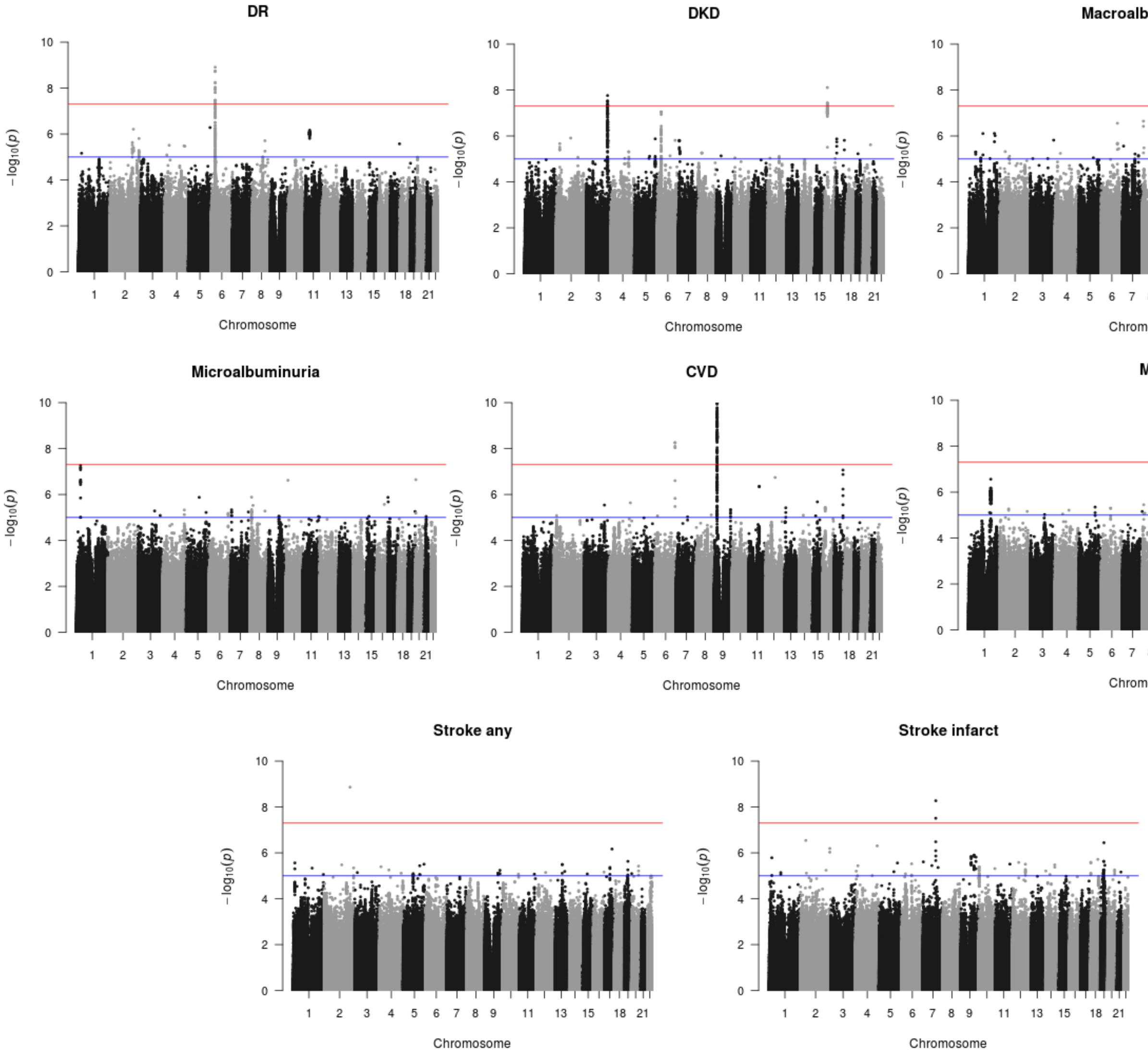

6291 Type II DM individuals with available genetic data
Excluded:
53 Estimated kinship >0.1768
Excluded:
12 Pre-imputation QC filter
6238 Unrelated type II DM individuals
6279 Type II DM individuals with no outlying missing rate
6226 Unrelated type II DM individuals with no outlying missing rate
Excluded:
1908 Self-identified non-whites
4318 NHW individuals with type II DM included in genetic analysis

Supplemental Figure 2. Diagram depicting a flow of participants used in the ACCORD analyses. DM, diabetes mellitus. NHW, none-Hispanic white.



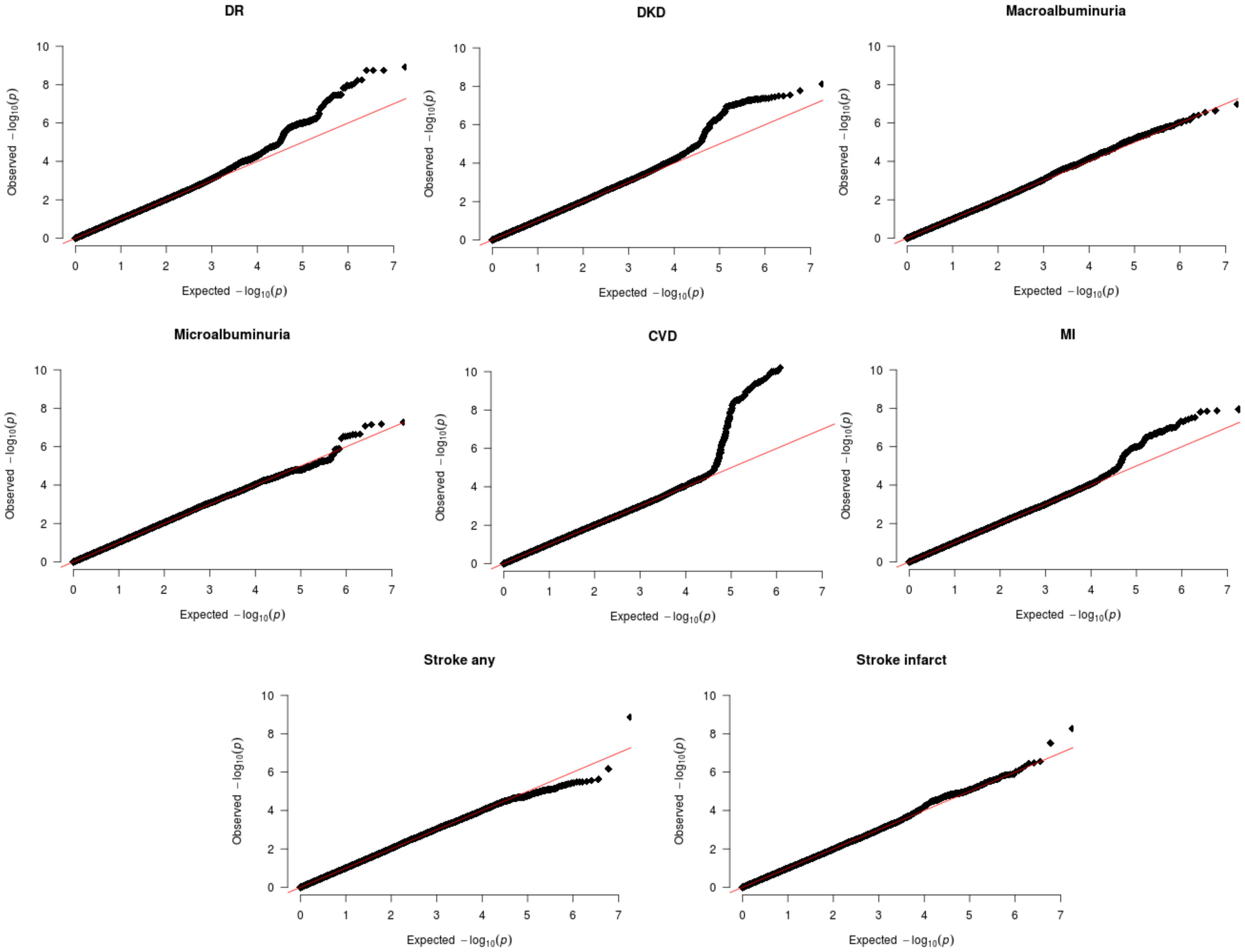
Supplemental Figure 3. Heritability estimates and standard errors of diabetes complication outcomes using the ACCORD genotype data and incorporating interaction with intensive glycemic treatment.

The grey bar represents the genetic plus interaction components, while the white bar signifies the interaction component. V(G), genetic variance. V(GxT), variance for interaction between genetics and intensive treatment. Vp, phenotypic variance.

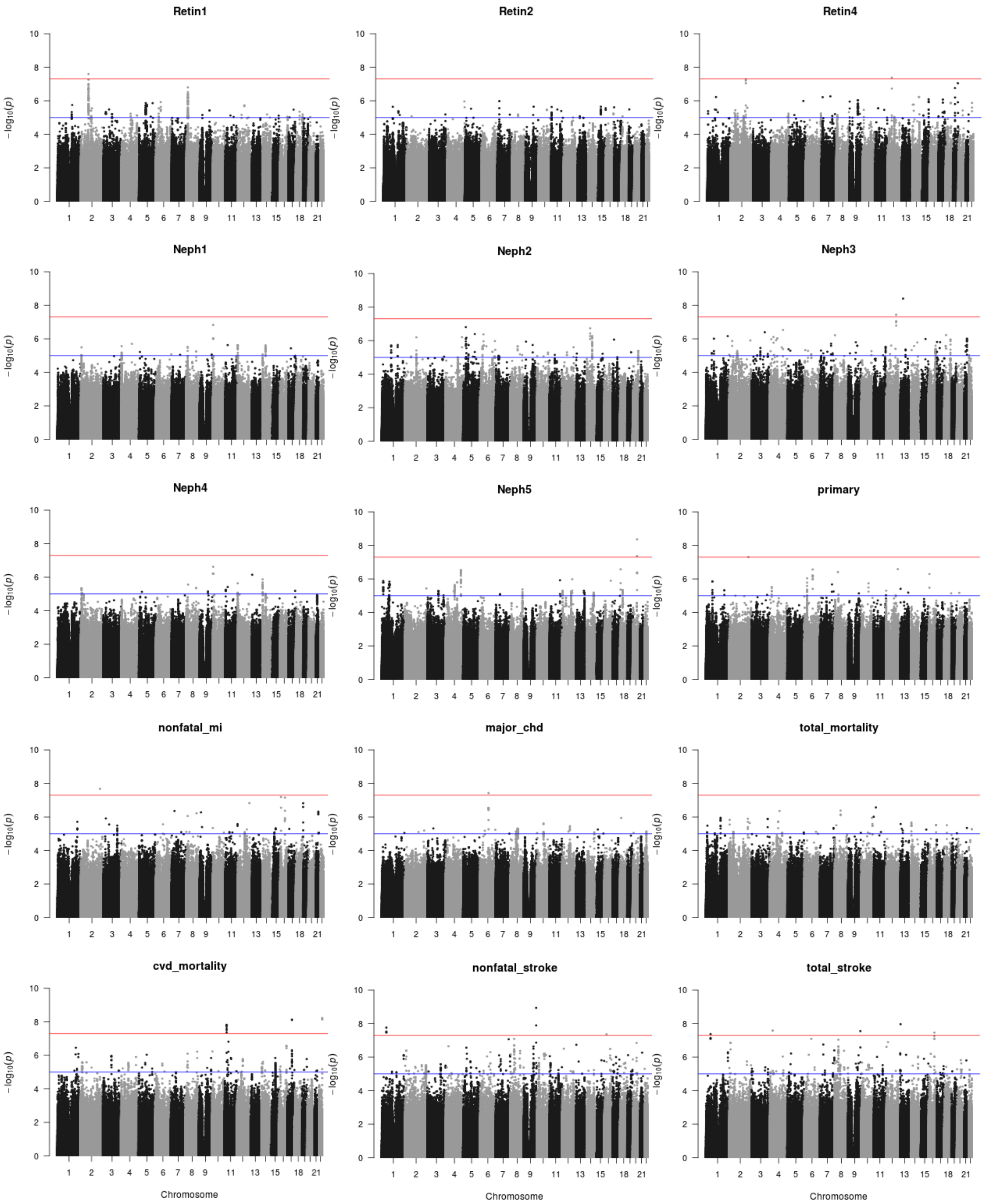


Supplemental Figure 4. Manhattan plots of GWAS *p*-values for the UKB phenotypes.

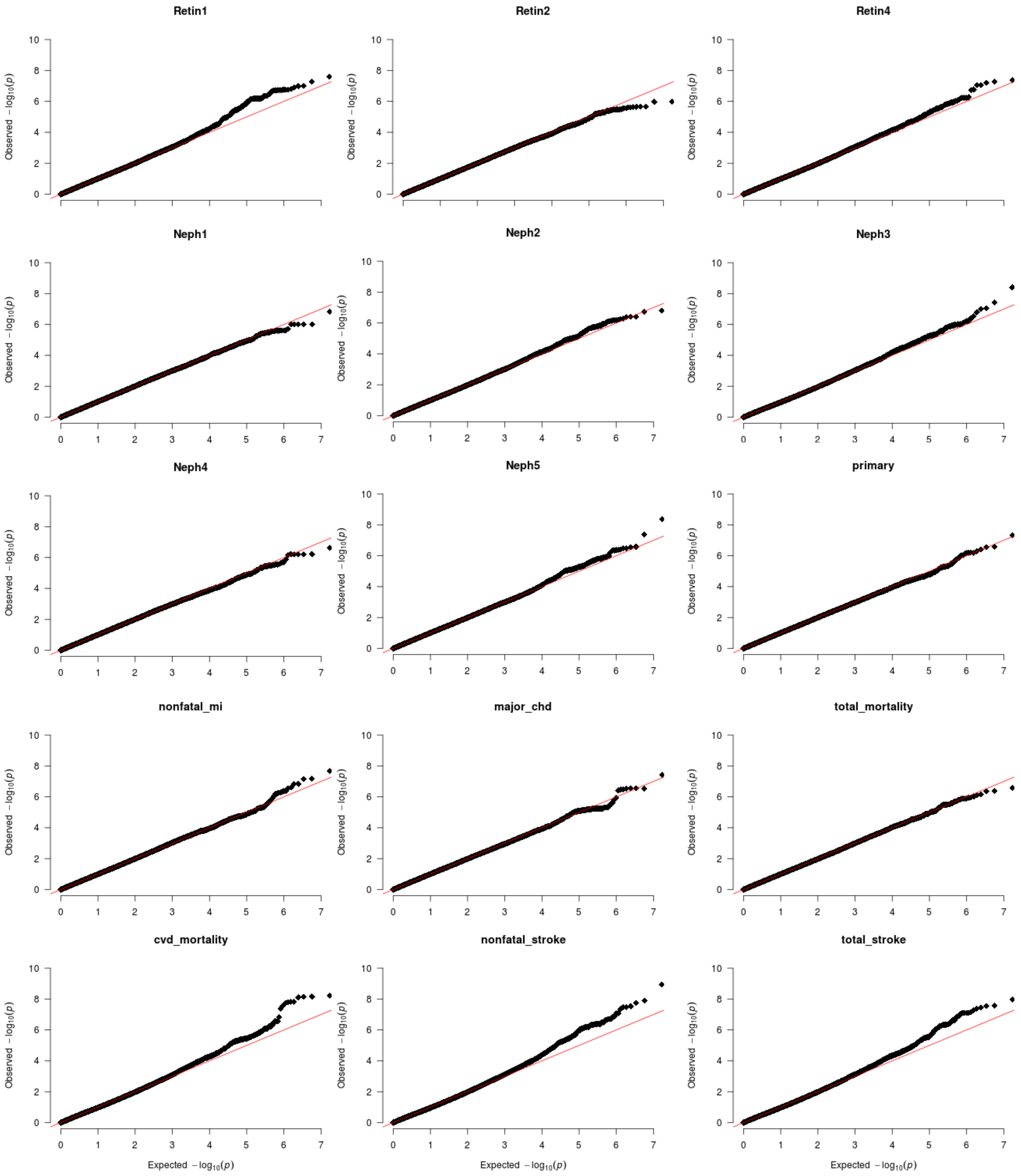
The red line signifies a genome-wide significance level ($p = 5\times10^{-8}$), while the blue line is a suggestive line ($p = 1\times10^{-5}$).



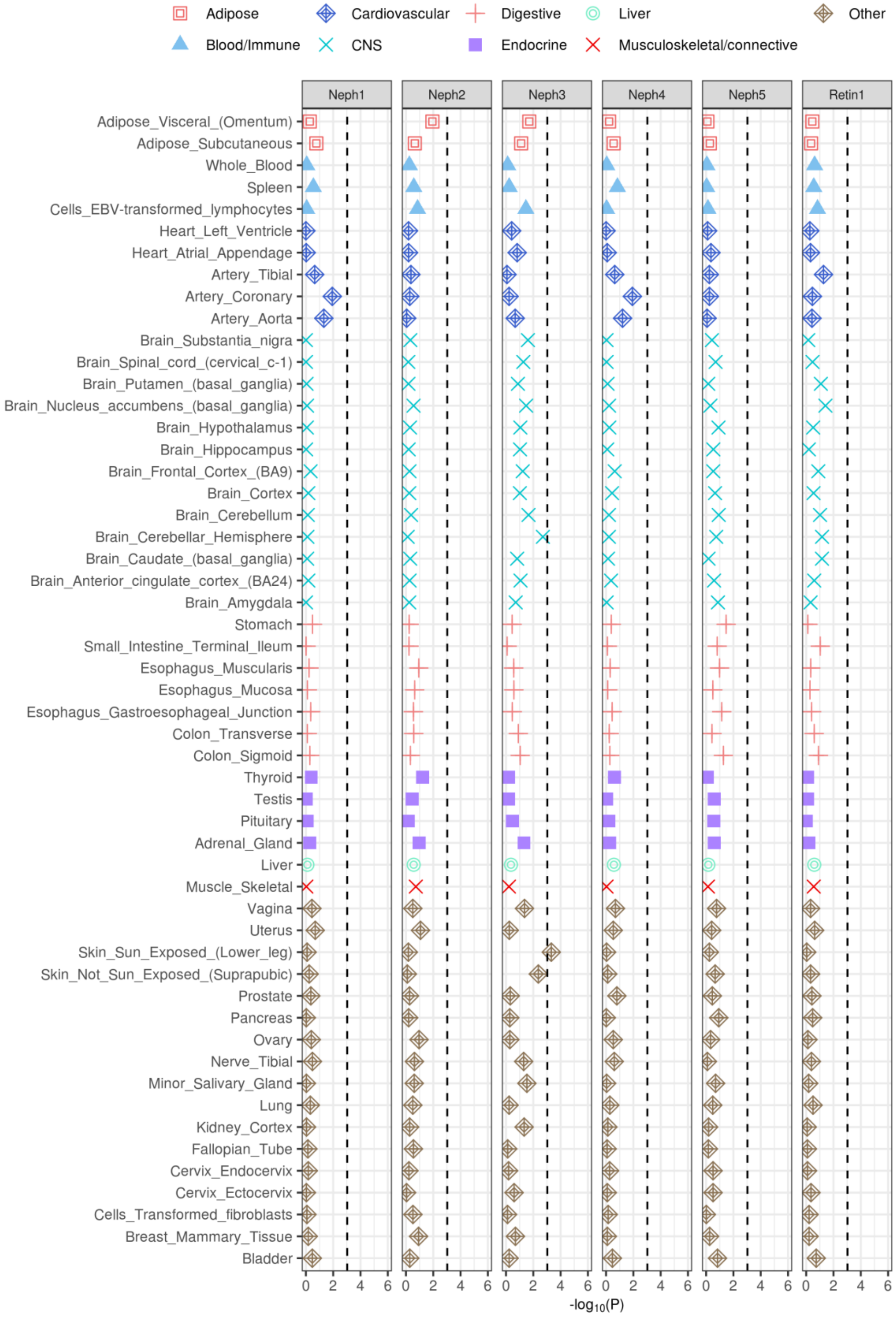
Supplemental Figure 5. QQ plots of GWAS *p*-values for the UKB phenotypes.



Supplemental Figure 6. Manhattan plots of GWAS *p*-values for the ACCORD phenotypes. The red line signifies a genome-wide significance level ($p = 5\times10^{-8}$), while the blue line is a suggestive line ($p = 1\times10^{-5}$).

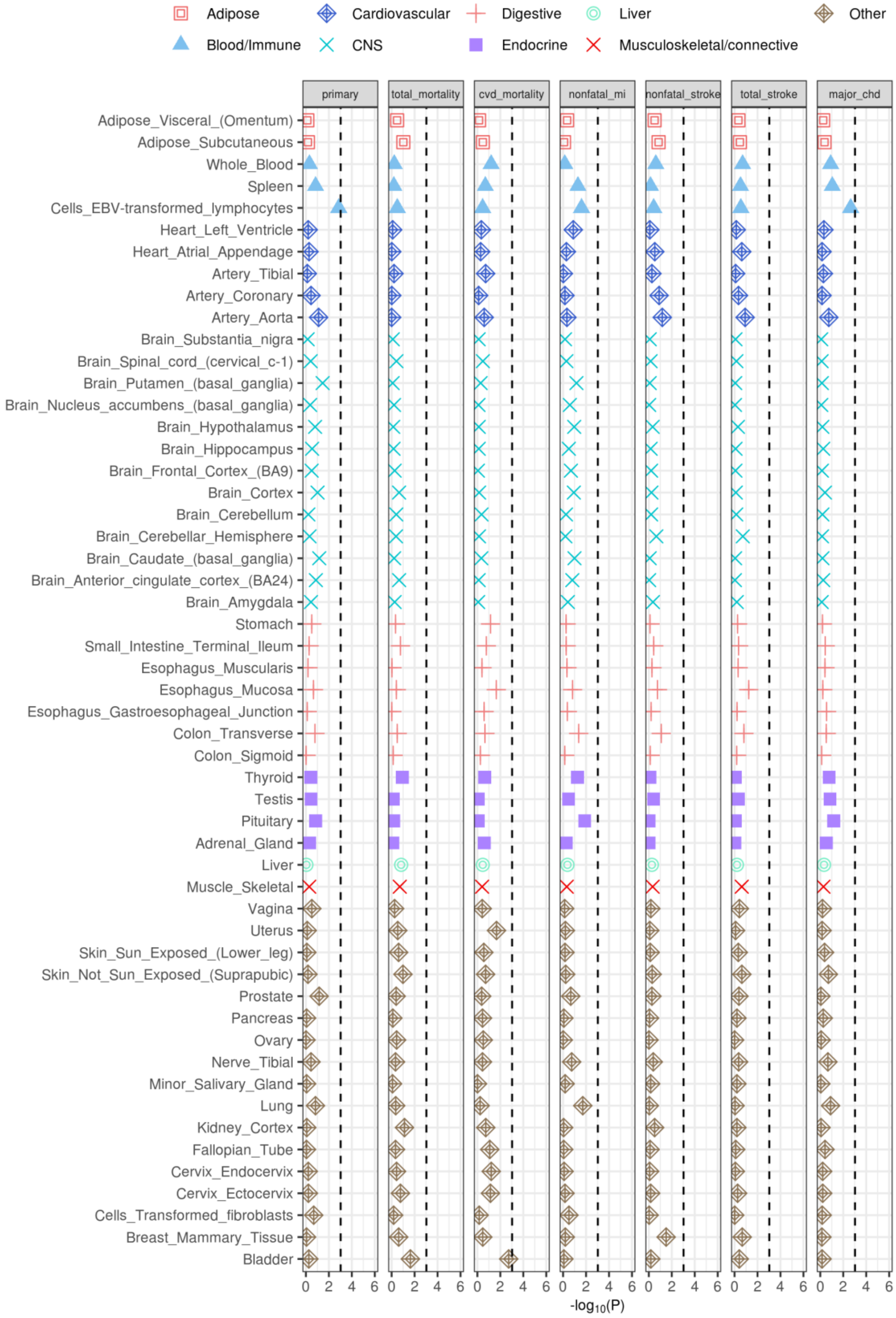


Supplemental Figure 7. QQ plots of GWAS *p*-values for the ACCORD phenotypes.



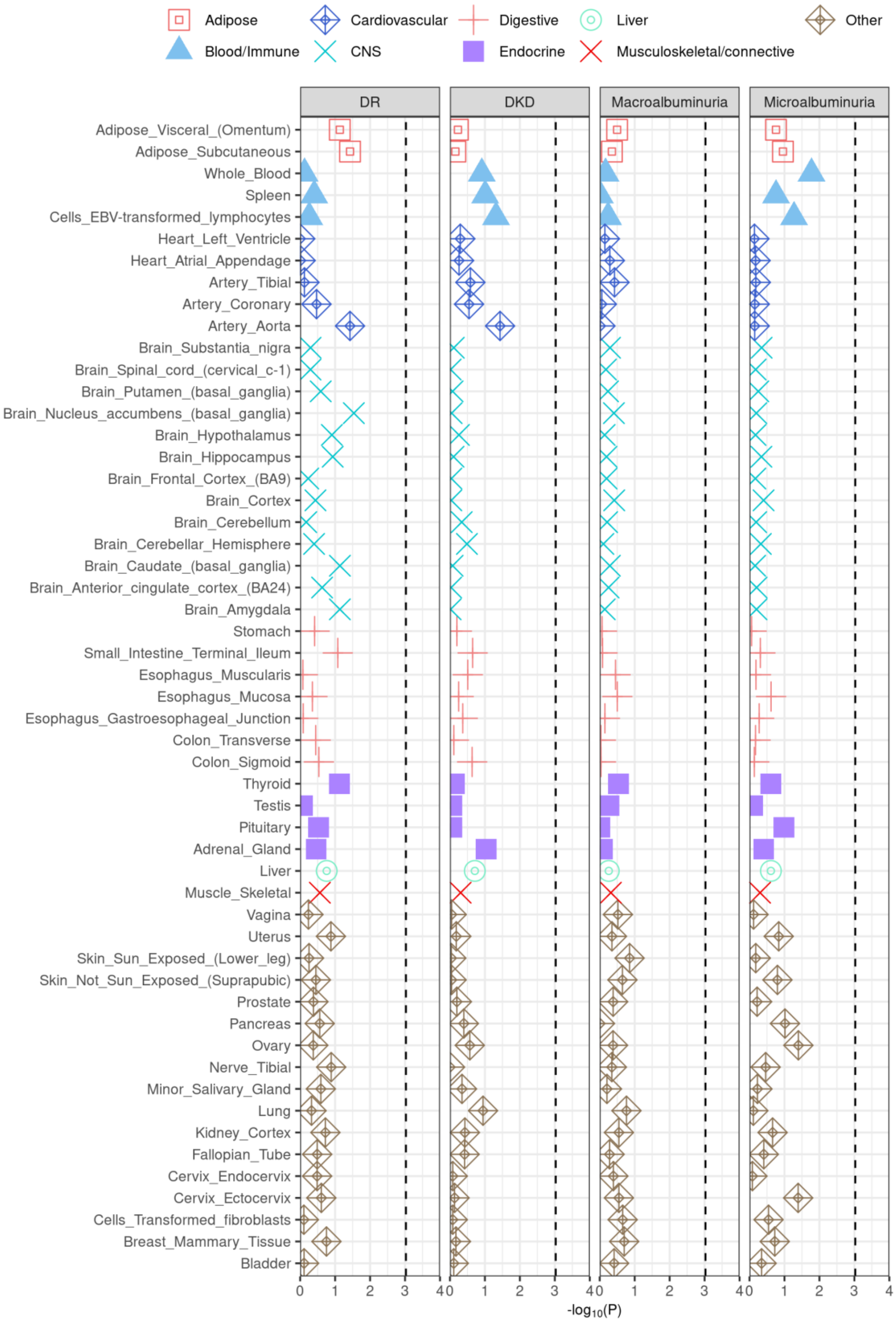
Supplemental Figure 8. Enrichment of the ACCORD microvascular complication phenotypes in tissue-specific gene expression annotations used in Finucane et al. (17).

The black dashed lines indicate the Bonferroni significance threshold ($p < 0.05/53$).



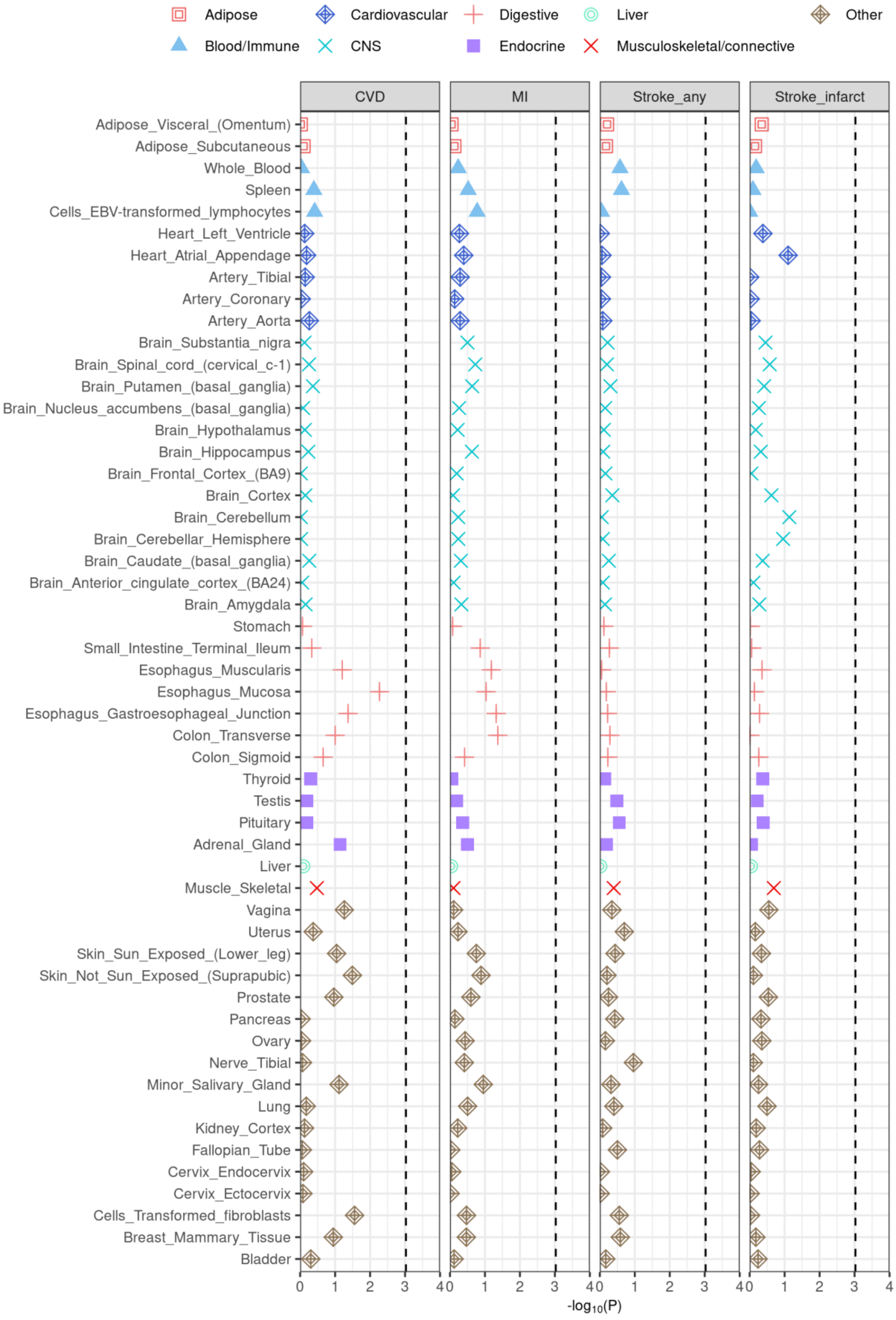
Supplemental Figure 9. Enrichment of the ACCORD macrovascular complication phenotypes in tissue-specific gene expression annotations used in Finucane et al. (1).

The black dashed lines indicate the Bonferroni significance threshold ($p < 0.05/53$).



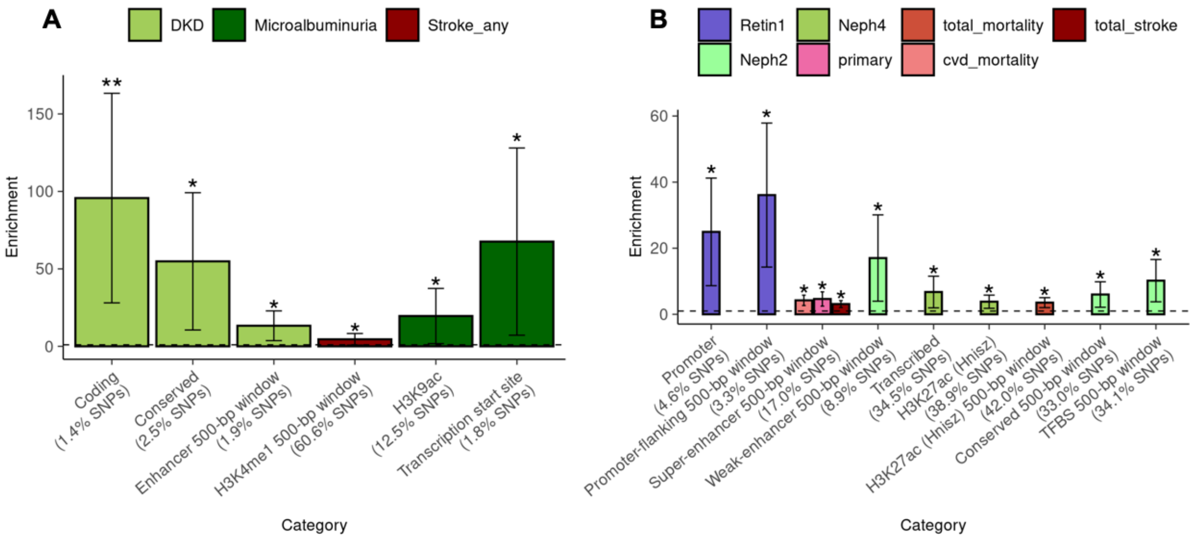
Supplemental Figure 10. Enrichment of the UKB microvascular complication phenotypes in tissue-specific gene expression annotations used in Finucane et al. (1).

The black dashed lines indicate the Bonferroni significance threshold ($p < 0.05/53$).



Supplemental Figure 11. Enrichment of the UKB macrovascular complication phenotypes in tissue-specific gene expression annotations used in Finucane et al. (1).

The black dashed lines indicate the Bonferroni significance threshold ($p < 0.05/53$).



Supplemental Figure 12. Enrichment estimates for selected annotations and traits using the (A) UKB and (B) ACCORD imputed data.

The dashed line represents no enrichment (enrichment=1). One asterisk indicates nominal significance at $p < 0.05$. Enrichment = $Pr(h^2)/Pr(SNPs)$. TFBS, Transcription factor binding site. DHS, DNase I hypersensitivity sites.

| | Neph1 | Neph2 | Neph4 | Neph5 | Retin1 |
|---|---|---|---|---|---|
| primary | -0.54 (0.41) | 0.12 (0.43) | -0.53 (0.40) | 0.14 (0.42) | -0.52 (0.39) |
| Neph1 | | 0.47 (0.62) | 0.96 (0.05) | 0.25 (0.57) | 0.19 (0.50) |
| Neph2 | | | 0.49 (0.57) | 0.11 (0.62) | 0.70 (0.67) |
| Neph4 | | | | 0.33 (0.56) | 0.32 (0.51) |
| Neph5 | | | | | 0.52 (0.59) |

Supplemental Table 1. Genetic correlation estimates and the standard errors between selected phenotypes using the ACCORD genotype data.

Adjusted for sex, CVD history at baseline, age at baseline, and the top five genetic principal components.

| | DKD | Microalbuminuria | DR |
|---|---|---|---|
| CVD | 0.25 (0.28) | -0.11 (0.24) | 0.26 (0.25) |
| DKD | | 0.36 (0.27) | 0.35 (0.35) |
| Microalbuminuria | | | 0.07 (0.25) |

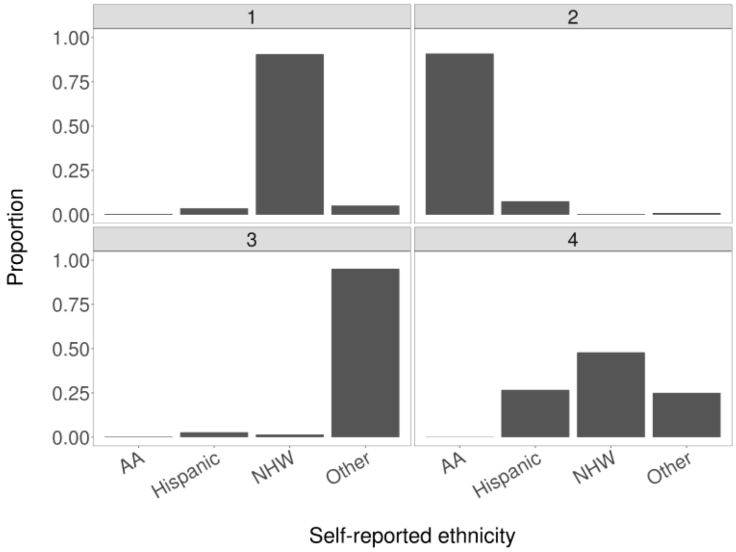
Supplemental Table 2. Genetic correlation estimates and the standard errors between selected phenotypes using the UKB genotype data.
Adjusted for sex, age in 2010, and the top ten genetic principal components.

## Methods

### Genotyping in UKB and ACCORD

**UKB**. Genome-wide genotyping was performed on all UK Biobank participants using the UK Biobank Axiom Array.

**ACCORD**. After downloading the data from dbGap (Study Accession: phs001411.v1.p1), we used genetic variants genotyped on Affymetrix Axiom Biobank 1 chips from the University of North Carolina (UNC) and merged data under two different institutional review board (IRB) protocols—HMB-IRB (73941) and DS-CDKD-IRB (73944). There were 6,291 (2,335 females and 3,956 males) with 546,800 SNPs in the merged dataset. Based on self-reported ethnicity, there were 4,369 non-Hispanic whites (NHW), 935 African-Americans (AA), 381 Hispanics, and 606 others. We checked the validity of self-reported ethnicity by running the ADMIXTURE software (2) with K=4, categorizing each individual into a group with the highest probability, and comparing the categories against self-reported ethnicity (see Supplemental Figure **13**). We can infer that the ADMIXTURE ancestry groups 1, 2, 3, and 4 represent NHW, AA, Other, and Hispanic, respectively. Considering that Hispanics are a highly genetically heterogeneous admixed group, the distribution in ADMIXTURE ancestry group 4 (Supplemental Figure **13**) appears reasonable.



Supplemental Figure 13. Bar graph indicating the percentage of self-reported ethnicity groups categorized into each ADMIXTURE bin.

Each individual is binned based on the largest proportion from ADMIXTURE.

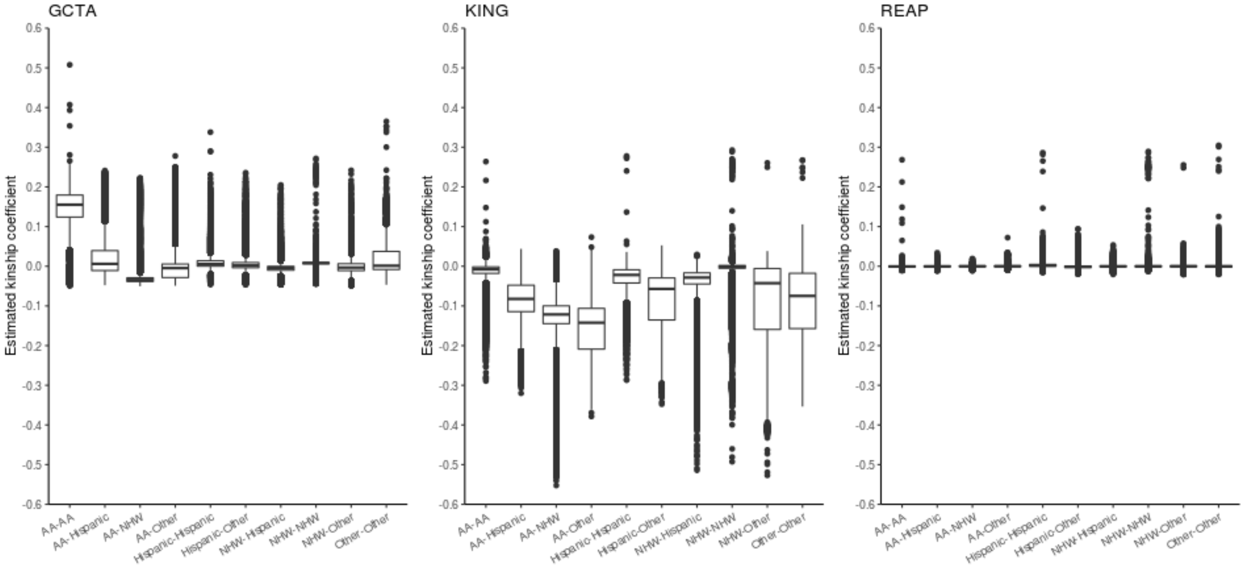
## Heritability estimation using genotype data

**UKB**. We extracted the NHW diabetes cohort (n=26,387) and computed the GRM via the REAP approach, for which necessary proportions were obtained from the ADMIXTURE software with K=3. No individuals were pruned out under the relatedness threshold (0.1768). We estimated heritability using the GREML-SC approach while adjusting for sex, age in 2010, and the top ten genetic principal components. Also calculated using the UKB genotype data were genetic correlations between phenotypes (see Supplemental Table **2**).

**ACCORD**. We calculated a Genetic Relationship Matrix (GRM) using SNPs from all autosomes. The GRM uses SNP data to measure the relatedness between each pair of individuals in our sample. This GRM replaces the known information about relatedness found in pedigrees. While the ACCORD trial did not deliberately recruit related individuals, we took a step to avoid inflation caused by cryptic (i.e., unknown) relatedness. We selectively excluded one of any pair of individuals with an estimated kinship greater than the separation between full and half-siblings (estimated kinship $> (1/2)^{5/2} = 0.1768$) in a way to maximize the remaining sample size (3; 4). Initially, we used the software package Genome-wide Complex Trait Analysis (GCTA) (5) to construct the GRM. However, the degree of relatedness calculated by GCTA appeared inflated (See Supplemental Figure **14**). The inflation may be mainly due to population heterogeneity in the data. Next, we tried Kinship-based INference for Genome-wide association studies (KING) (3). As seen in Supplemental Figure **14**, estimated kinship-coefficient values from KING were systematically negative, which ultimately led the GRM to be not positive semi-definite. Finally, we used Relatedness Estimation in Admixed Populations (REAP) (6), which produced more robust results. The REAP approach requires individual ancestry proportions and allele frequencies for each ancestral population. Both proportions were obtained using the

ADMIXTURE software (2) with the number of ancestral populations specified as four (K=4). The number four was chosen because there were four different self-reported ethnic groups (NHW, AA, Hispanic and other).

We only extracted NHW samples after pruning related individuals, leaving us with 4,329 samples. With the GRM constructed from REAP, heritability was estimated via GCTA (4). We adjusted for sex, CVD history at baseline, age at baseline, and the top five genetic principal components. An additional analysis that incorporated interaction with glycemic intensive treatment arm (intensive=1, standard=0) is shown in Supplementary Figure 3. We also estimated the genetic correlation between binary traits via the GCTA software (4; 8), including sex, CVD history at baseline, age at baseline, and the top five genetic principal components as covariates.



Supplemental Figure 14. Estimated kinship coefficients from software packages GCTA, KING, and REAP.

Estimates from GCTA have been divided by 2 for comparability with other packages.

## Imputation

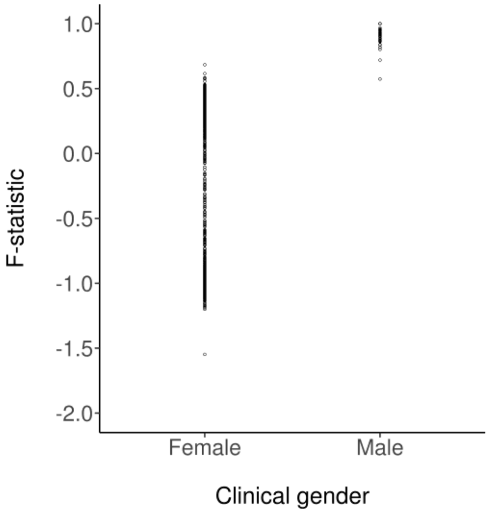
**UKB**. We used the imputed datasets released by UK Biobank. After extracting autosomal variants with imputation info score > 0.3 and removing multiallelic variants from the imputed datasets, we excluded variants with missing genotype rate > 0.05, HWE test $p < 1\times10^{-6}$, and MAF < 0.0001. After the filtering steps, we had a total of 33,932,888 variants.

**ACCORD**. Prior to imputation, we performed quality control steps on the data. First, we checked if there are mismatches between genetic gender and clinical gender. We ran plinkv1.9 --check-sex option along with --split-x and made an F-statistic against sex-label plot (see Supplemental Figure 15). As expected, we saw a big tight clump near 1 for males while a more widely dispersed set of values centered near 0 (7). Even though some individuals did not pass the default threshold set in plink, we decided not to remove any individuals since the data exhibit an expected pattern.

Next, following the procedure in (4), we computed Hardy-Weinberg equilibrium (HWE) $\chi^2$ values for each of the self-reported ethnicity groups: NHW, AA, Hispanic, and other. Any SNPs deviating from $p$ value $1\times10^{-5}$ in at least two of the four groups were excluded. This step reduced the number of variants to 542,847. Additionally, we checked alleles to allow only A, C, G, T and excluded SNPs with a missing rate > 3% and monomorphic sites (MAF < 0.0000001). We also excluded individuals with a genotype missing rate > 0.03. After the aforementioned step, we retained 6,279 individuals and 465,011 variants.

Data imputation was done using a two-step approach where the genotype calls were pre-phased using Eagle v2.4.1 (8) and then imputation was done using Minimac4 (9) with default options. Both steps used the 1000 Genomes Project Phase 3 (10) as a reference panel.

After discarding imputed variants with $R^2 < 0.3$ and MAF < 0.0003, we had a total of 25,667,109 imputed variants for the downstream analyses. Additionally, we extracted the NHW samples filtered from the REAP approach earlier (n=4,329). With 11 out of 4,329 individuals removed during the pre-imputation QC steps, we proceeded with the downstream analyses with 4,318 NHW individuals.



Supplemental Figure 15. Distribution of F (inbreeding) coefficients against clinical gender.

### GREML-LDMS

On the imputed datasets, we employed the GREML-LDMS method. For the GREML-LDMS-I approach, we followed the design laid out in (11). First, we calculated segment-based LD scores using the default settings—200-kb block size with a 100-kb overlap—using the GCTA software and stratify SNPs into high LD and low LD score groups using the median as a threshold. In each LD group, SNPs were further partitioned into four MAF bins: common ($MAF \geq 0.05$), uncommon ($0.01 \leq MAF < 0.05$), rare ($0.0025 \leq MAF < 0.01$), and very rare ($0.0003 \leq MAF < 0.0025$). Then GRMs were computed using SNPs stratified into eight groups, hence creating eight GRMs. Finally, we ran GREML analyses on each binary phenotype with fixed covariates.

**UKB**. On the UKB imputed datasets, we adjusted for sex, age in 2010, and the top ten genetic principal components.

**ACCORD**. After filtering steps, the ACCORD imputed dataset contained 4,318 NHW individuals and 15,349,988 variants. We included sex, age at baseline, history of CVD at baseline, and the top five genetic principal components as covariates.

## GWAS

**UKB**. GWAS for complications was performed in 26,387 NHW samples. After MAF filtration (MAF ≥ 0.01), 8,949,996 variants formed the GWAS panel. We adjusted for sex, age in 2010, and the top ten genetic principal components. Manhattan and QQ plots are provided in Supplemental Figure 4 and Supplemental Figure 5 and respectively.

**ACCORD**. GWAS for complications were performed in 4,318 NHW participants. After filtration for variants with MAF ≥ 0.01, as done in Bulik-Sullivan et al. (12), 8,480,081 SNPs formed the GWAS panel. The association between each variant and each complication was tested by logistic regression in PLINK2.0 (7), assuming an additive genetic model and adjusting for sex, CVD history at baseline, age at baseline, and the top five genetic principal components. Manhattan and quantile-quantile (QQ) plots are provided in Supplemental Figure 6 and Supplemental Figure 7, respectively.

## Stratified LD score regression (S-LDSC)

We partitioned SNP heritability, applying S-LDSC to GWAS summary statistics for the trait of interest. We conducted S-LDSC analysis using the ‘full baseline model’ generated by Finucane al. (13). The full baseline model is comprised of 53 overlapping functional categories (including coding, promoter, enhancer, and conserved regions) and is not specific to any cell type. We also conducted tissue-type specific analyses where we used the 53 specifically expressed gene annotations curated from the Genotype-Tissue Expression (GTEx) project (14) by Finucane et al. (1). For all S-LDSC analyses, we used 1000 Genomes Project Phase 3 (10) European population SNPs as an LD reference panel. All annotations and reference panel data were obtained from Alkes Price’s group data repository (see URLs).

## URLs

Baseline LDSC annotations, https://data.broadinstitute.org/alkesgroup/LDSCORE/; Finucane GTEx annotations, https://data.broadinstitute.org/alkesgroup/LDSCORE/LDSC_SEG_ldscores/; LDSC, https://github.com/bulik/ldsc/wiki.

# References

1. Finucane HK, Reshef YA, Anttila V, Slowikowski K, Gusev A, Byrnes A, Gazal S, Loh P-R, Lareau C, Shoresh N: Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat Genet* 2018;50:621-629

2. Alexander DH, Novembre J, Lange K: Fast model-based estimation of ancestry in unrelated individuals. Genome Res 2009;19:1655-1664

3. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M: Robust relationship inference in genome-wide association studies. *Bioinformatics* 2010;26:2867-2873

4. Marvel SW, Rotroff DM, Wagner MJ, Buse JB, Havener TM, McLeod HL, Motsinger-Reif AA: Common and rare genetic markers of lipid variation in subjects with type 2 diabetes from the ACCORD clinical trial. *PeerJ* 2017;5:e3187

5. Yang J, Lee SH, Goddard ME, Visscher PM: GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 2011;88:76-82

6. Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, Risch N: Estimating kinship in admixed populations. *Am J Hum Genet* 2012;91:122-138

7. Chang CC: Data Management and Summary Statistics with PLINK. *Methods Mol Biol* 2020;2090:49-65

8. Loh P-R, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, Finucane HK, Schoenherr S, Forer L, McCarthy S, Abecasis GR: Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* 2016;48:1443

9. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M: Next-generation genotype imputation service and methods. *Nat Genet* 2016;48:1284-1287

10. The 1000 Genomes Project Consortium: A global reference for human genetic variation. *Nature* 2015;526:68

11. Evans LM, Tahmasbi R, Vrieze SI, Abecasis GR, Das S, Gazal S, Bjelland DW, De Candia TR, Goddard ME, Neale BM: Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits. *Nat Genet* 2018;50:737-745

12. Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Patterson N, Daly MJ, Price AL, Neale BM: LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 2015;47:291-295

13. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, Anttila V, Xu H, Zang C, Farh K: Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* 2015;47:1228

14. GTEx Consortium: The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* 2015;348:648-660