

Common and Rare Variants Associated with Cardiometabolic Traits across 98,622 Whole-Genome Sequences in the *All of Us* Research Program

Supplemental Materials

Supplementary Author Information	2
Supplementary Figures	4
Figure S1. Manhattan plot of genome-wide association study of coronary artery disease	4
Figure S2. LocusZoom plot of significant risk locus in the atrial fibrillation GWAS	5
Figure S3. LocusZoom plot of significant risk loci in the type 2 diabetes GWAS	6
Figure S4. LocusZoom plot of significant loci in the low-density lipoprotein GWAS	7
Supplementary Tables	9
Table S1. Phenotype definition for atrial fibrillation	9
Table S2. Summary statistics of the lead common variants identified in the <i>All of Us</i> GWAS and in prior GWAS	11
Supplementary Methods	13
Study population	13
Cardiometabolic phenotypes	13
Whole-Genome Sequencing and variant calling	14
Variant-level and sample-level quality control	15
Principal components	15
Common variant analysis and genetic correlation	16
Rare variant analysis using burden test	17
References	19

Supplementary Author Information

Flagship Manuscript Writing Group: Alexander Bick MD, PhD¹, Ginger A Metcalf BS⁹, Ashley Able PhD², Kelsey R Mayo PhD², Lee Lichtenstein MS⁵, Shimon Rura BA⁷, Robert Carroll PhD³, Anjene Musick PhD, MPH¹⁵

All of Us Genomic PIs: Eric Boerwinkle PhD⁹, Mine S Cicek PhD¹³, Kimberly F Doheny PhD¹⁰, Evan E Eichler PhD¹², Stacey Gabriel PhD⁸, Richard A Gibbs PhD⁹, David Glazer BS⁶, Paul Harris PhD³, Gail P Jarvik MD, PhD¹¹, Anthony Philippakis MD, PhD⁵, Heidi L Rehm PhD⁸, Dan Roden MD⁴, Stephen N Thibodeau PhD¹³, Scott Topper PhD⁶

Biobank, Mayo: Ashley L Blegen¹⁴, Samantha J Wirkus¹⁴, Jeffrey G Meyer¹⁴, Mine S Cicek PhD¹³, Stephen N Thibodeau PhD¹³,

Data and Research Center, Broad Institute of MIT & Harvard: Lee Lichtenstein MS⁵, Sophie Schwartz PhD⁵, M. Morgan Taylor PhD⁵, Kristian Cibulskis⁵, Andrea Haessly MS⁵, Rebecca Asch BLA⁵, Aurora Cremer⁵, Kylee Degatano MS⁵, Akum Shergill⁶, Laura Gauthier PhD⁶, Eric Banks PhD⁵, Anthony Philippakis MD, PhD⁵

Data and Research Center, Vanderbilt University Medical Center: Melissa Basford MBA², Alexander Bick MD, PhD¹, Ashley Able PhD², Kelsey R Mayo PhD², Robert Carroll PhD³, Jennifer Zhang MS², Henry Condon¹, Yuanyuan Wang PhD², Paul Harris PhD³, Dan Roden MD⁴

Data and Research Center, Verily: Shimon Rura BA⁷, Moira K Dillon BS⁷, CH Albach BS⁷, David Glazer BS⁶

Genome Center, Baylor College of Medicine: Richard A Gibbs PhD⁹, Eric Boerwinkle PhD⁹, Donna M Muzny MS⁹, Ginger A Metcalf BS⁹, Eric Venner PhD⁹, Kimberly Walker MS⁹, Jianhong Hu PhD⁹, Harsha Doddapaneni PhD⁹, Christie L Kovar BS⁹, Mullai Murugan BS⁹, Shannon Dugan BS⁹, Ziad Khan BS⁹

Genome Center, Broad Institute of MIT & Harvard: Stacey Gabriel PhD⁸, Heidi L Rehm PhD⁸, Scott Topper PhD⁶, Niall J Lennon PhD⁸, Namrata Gupta PhD⁸, Alicia Zhou PhD⁶, Cynthia Neben PhD⁶, Christopher Kachulis PhD⁸

Genome Center, Johns Hopkins University School of Medicine: Kimberly F Doheny PhD¹⁰, Michelle Z Mawhinney MS¹⁰, Sean ML Griffith MS¹⁰, Elvin Hsu BS¹⁰, Hua Ling PhD¹⁰, Marcia K Adams MS¹⁰

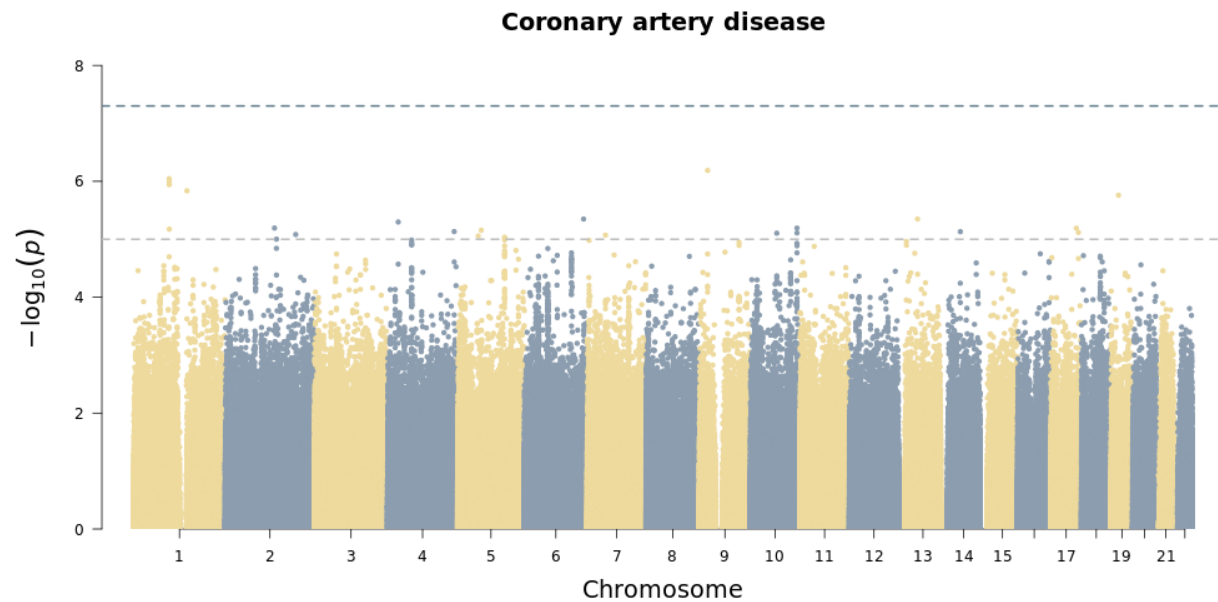
Genome Center, University of Washington School of Medicine: Gail P Jarvik MD, PhD¹¹, Evan E Eichler PhD¹², Joshua D Smith MS¹², Christian D Frazar MS¹², Colleen P Davis BS¹², Karynne E Patterson BS¹², Marsha M Wheeler PhD¹², Sean McGee PhD¹², Aparna Radhakrishnan PhD¹²

NIH *All of Us* Research Program Staff: Andrea H Ramirez MD, MS¹⁵, Sokny Lim MPH¹⁵, Siddhartha Nambiar PhD¹⁵, Anjene Musick PhD, MPH¹⁵, Bradley Ozenberger PhD¹⁵, Chris Lunt BS¹⁵, Joshua Denny MD, MS¹⁵

¹Department of Genetic Medicine, Vanderbilt University Medical Center, Nashville, TN, 37203, USA, ²Vanderbilt Institute of Clinical and Translational Research, Vanderbilt University Medical Center, Nashville, TN, 37203, USA, ³Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, 37203, USA, ⁴Department of Clinical Pharmacology, Vanderbilt University Medical Center, Nashville, TN, 37203, USA, ⁵Broad Institute of MIT and Harvard, Data Sciences Platform, Cambridge, MA, 02142, USA, ⁶ Broad Institute of MIT and Harvard, Data Sciences Platform, Cambridge, MA, 02142, USA, ⁷Verily Life Sciences, South San Francisco, CA, USA, 94080, ⁸Broad Institute or MIT and Harvard, Cambridge, MA 02142 USA, ⁹Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX , 77030 USA, ¹⁰Department of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA, ¹¹Department of Medicine, Division of Medical Genetics, University of Washington School of Medicine, Seattle, WA 98195 USA, ¹²Department of Genome Sciences, University of Washington School of Medicine, Seattle, WA 98195, USA, ¹³Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN 55905, USA, ¹⁴Center for Individualized Medicine, Biorepository Program, Mayo Clinic, Rochester, MN 55905, USA, ¹⁵All of Us Research Program, National Institutes of Health, Bethesda, MD 20817

Supplementary Figures

Figure S1. Manhattan plot of genome-wide association study of coronary artery disease



Chromosomal variant positions are plotted on the x-axis. The $-\log_{10}(P)$ values are plotted on the y-axis. The genome-wide significance threshold (5×10^{-8}) is indicated by the horizontal dotted line.

Figure S2. LocusZoom plot of significant risk locus in the atrial fibrillation GWAS

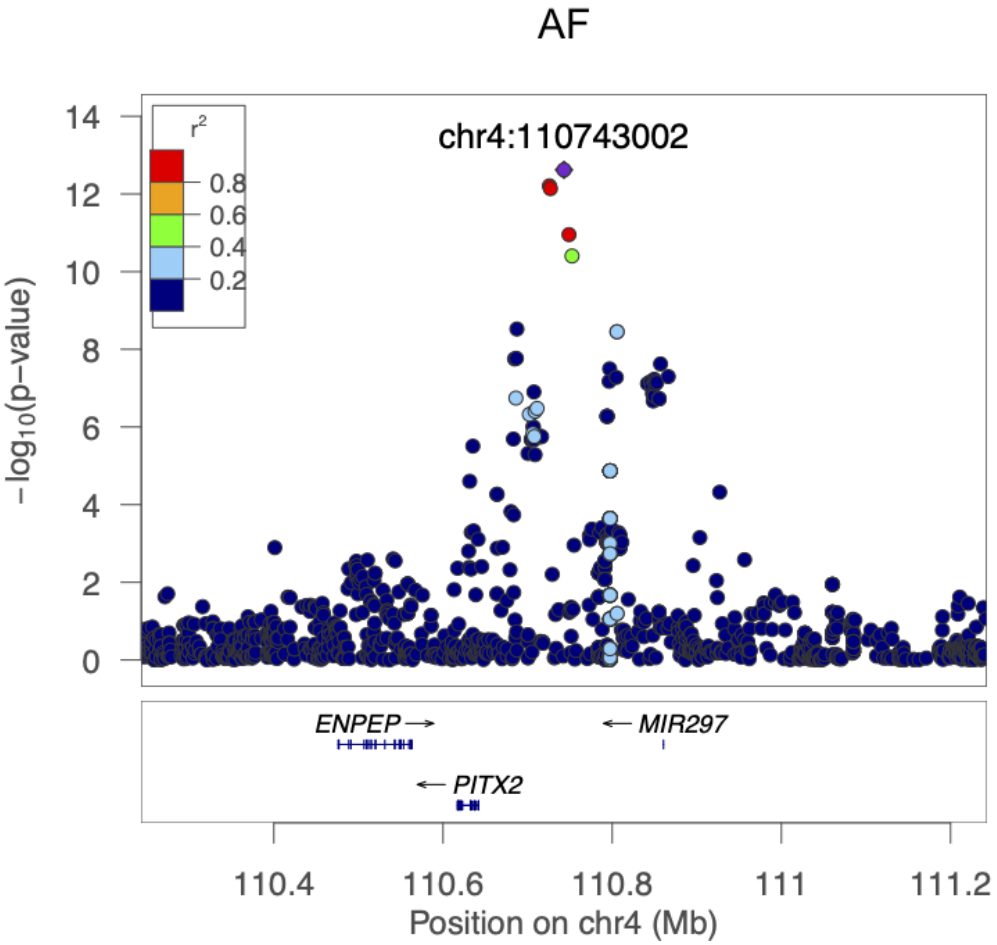


Figure S3. LocusZoom plot of significant risk loci in the type 2 diabetes GWAS

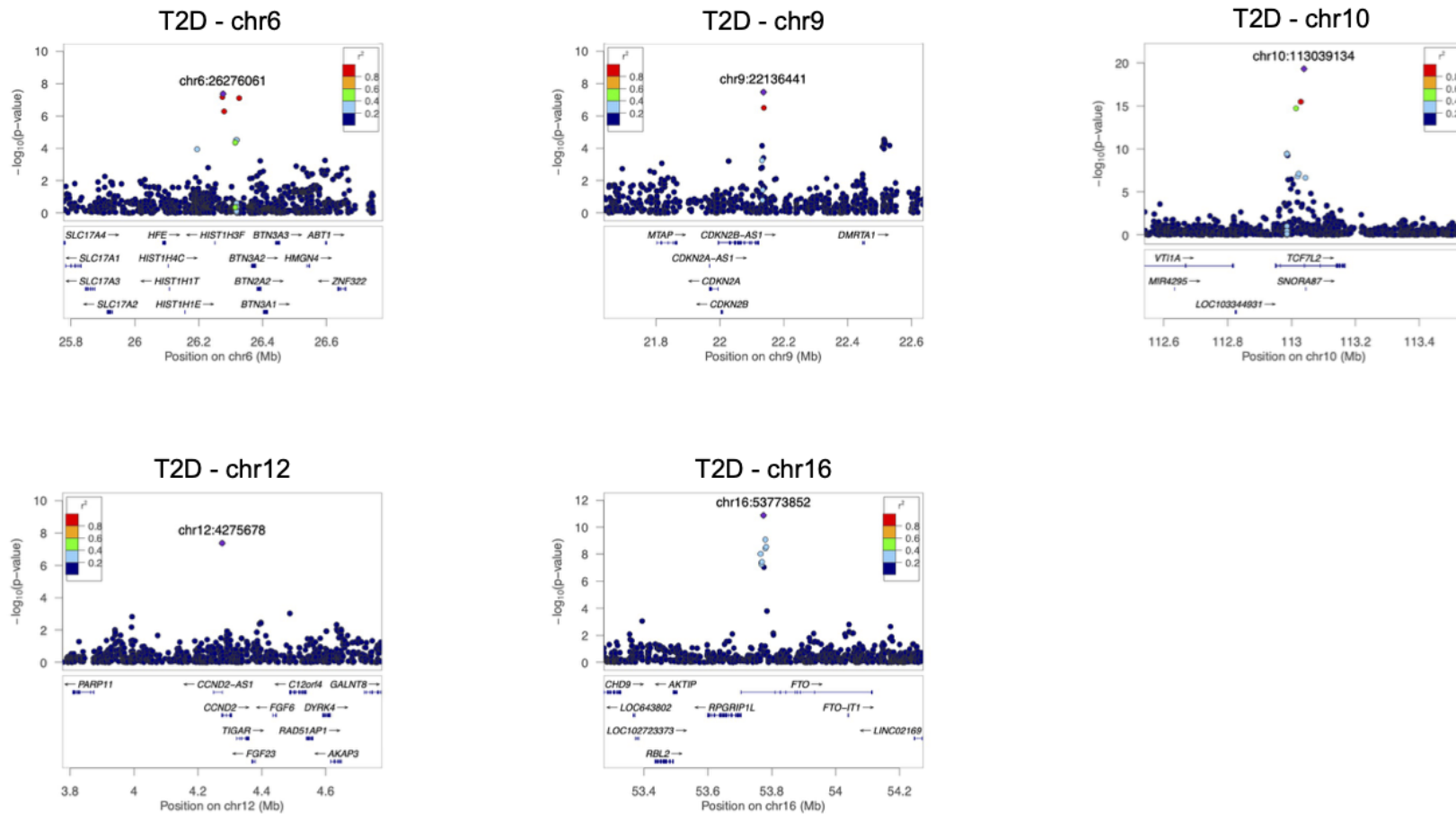
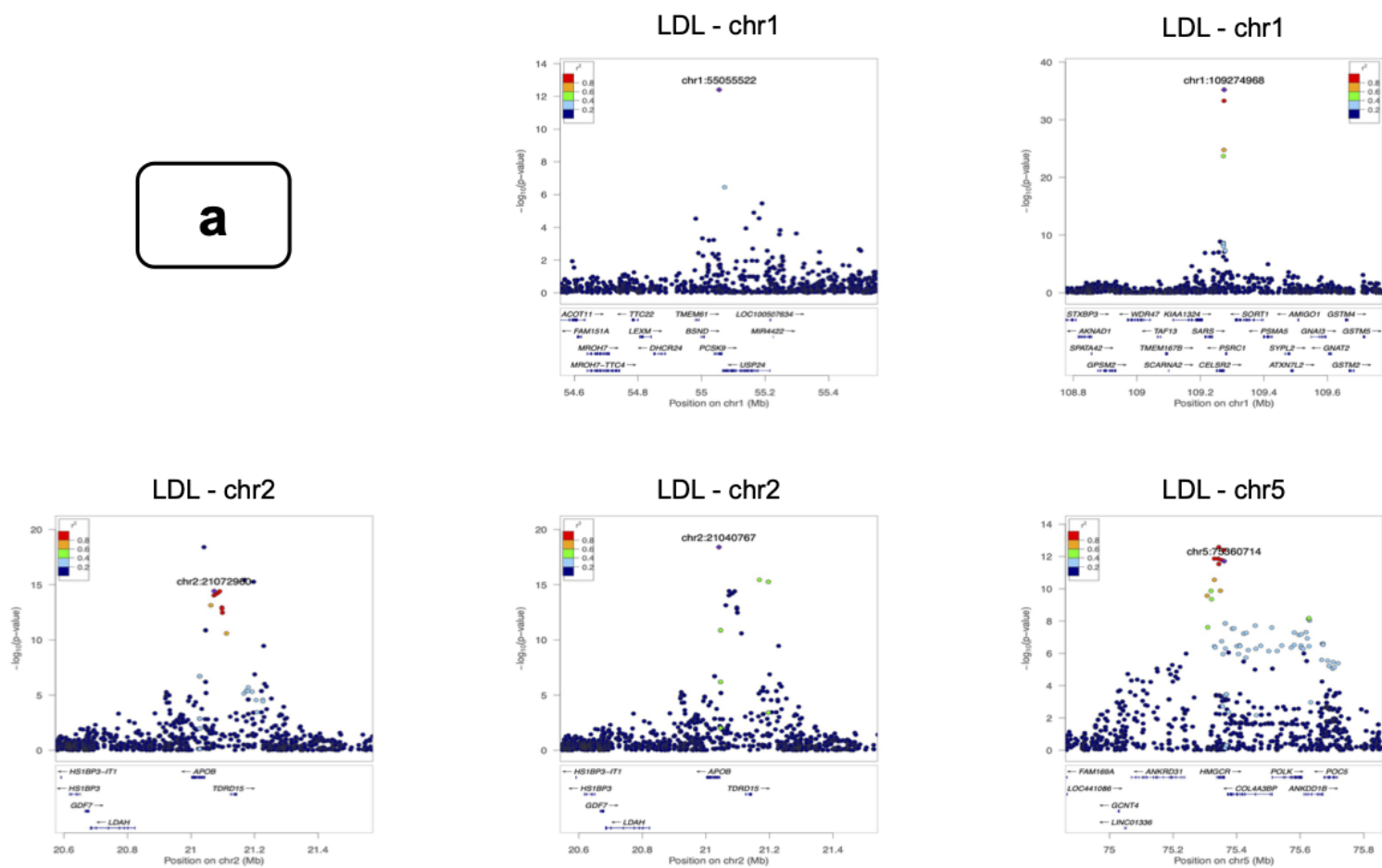
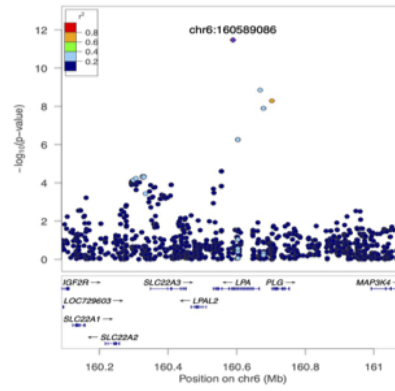


Figure S4. LocusZoom plot of significant loci in the low-density lipoprotein GWAS

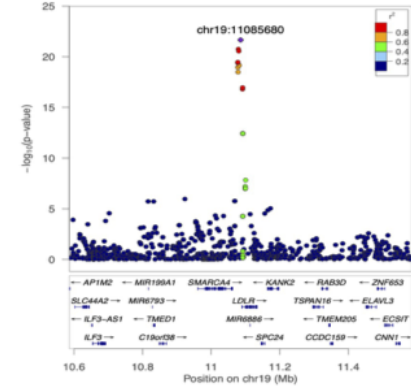


b

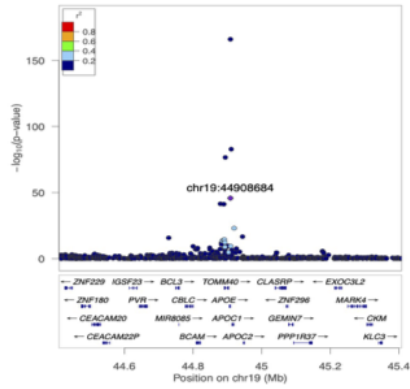
LDL - chr6



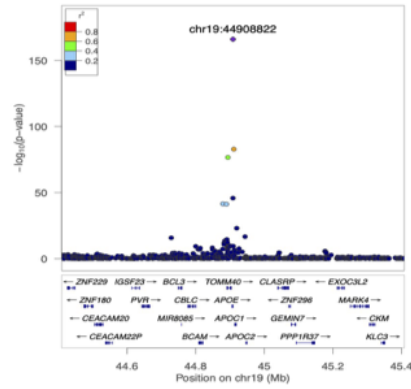
LDL - chr19



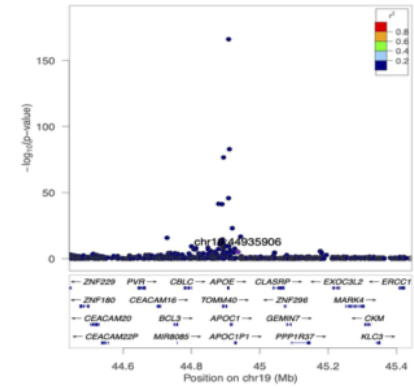
LDL - chr19



LDL - chr19



LDL - chr19



Supplementary Tables

Table S1. Phenotype definition for atrial fibrillation

ICD codes	ICD 9: 427.3, 427.31, 427.32 ICD 10: I48, I48.0, I48.1, I48.2, I48.3, I48.4, I48.9
Self-reported medical history	Circulatory Conditions: Atrial Fibrillation (Concept id: 43528442)
Procedures/Operations	<ol style="list-style-type: none"> 1. Cardioversion, elective, electrical conversion of arrhythmia; external (CPT4 Code: 92960; Concept Id: 2313791) 2. Cardioversion (SNOMED Code: 250980009; Concept Id: 4353741) 3. Atrial cardioversion (SNOMED Code: 26879000; Concept Id: 4098410) 4. External electrode cardioversion (SNOMED Code: 275148007; Concept Id: 4166447) 5. Cardioversion, elective, electrical conversion of arrhythmia; internal (separate procedure) (CPT4 Code: 92961, Concept Id: 2313792) 6. Comprehensive electrophysiologic evaluation including transeptal catheterizations, insertion and repositioning of multiple electrode catheters with induction or attempted induction of an arrhythmia including left or right atrial pacing/recording when nec

	(CPT4 Code: 93656; Concept Id: 43528008)
--	---

Concept IDs are codes specific to the *All of Us* database and can be used to query data.

Participants with any of these codes in their electronic health records (EHR) were identified as cases for atrial fibrillation.

Table S2. Summary statistics of the lead common variants identified in the *All of Us* GWAS and in prior GWAS

Genetic variants in the AoU GWAS [ref/effect] (GRCh38)	Effect size [95% CI] in the AoU GWAS	P-value in the AoU GWAS	Genetic variants in prior GWAS [ref/effect] (GRCh37)	Effect size [95% CI] in prior GWAS	P-value in the GWAS
Atrial Fibrillation					
chr4:110743002 [G/A]	1.26 [1.18, 1.34]	2.40×10^{-13}	chr4:111664158 [G/A]	1.56 [1.53, 1.59]	8.80×10^{-453}
Type 2 Diabetes					
chr6:26276061 [T/G]	0.79 [0.73, 0.86]	4.22×10^{-8}	Not in prior GWAS		
chr9:22136441 [G/C]	1.11 [1.07, 1.16]	3.39×10^{-8}	Not in prior GWAS		
chr10:113039134 [T/A]	1.19 [1.15, 1.24]	4.87×10^{-20}	Not in prior GWAS		
chr12:4275678 [T/G]	0.66 [0.57, 0.77]	4.23×10^{-8}	chr12:4384844 [T/G]	0.66 [0.63, 0.68]	1.20×10^{-96}
chr16:53773852 [A/G]	1.12 [1.08, 1.16]	1.33×10^{-11}	chr16:53807764 [A/G]	1.08 [1.05, 1.10]	2.70×10^{-9}
Low-density lipoprotein					
chr1:55055522 [C/T]	-0.09 [-0.11, -0.06]	1.71×10^{-12}	chr1:55521195 [C/T]	-0.08 [-0.06, -0.09]	3.70×10^{-20}

chr1:109274968 [G/T]	-0.12 [-0.13, -0.10]	3.49×10^{-38}	chr1:109817590 [G/T]	-0.12 [-0.12, -0.11]	9.58×10^{-411}
chr2:21040767 [T/G]	0.10 [0.08, 0.12]	1.59×10^{-21}	chr2:21263639 [T/G]	0.10 [0.09, 0.11]	3.53×10^{-195}
chr2:21072960 [G/A]	0.07 [0.06, 0.09]	7.75×10^{-20}	chr2:21295832 [G/A]	0.06 [0.06, 0.07]	1.34×10^{-174}
chr5:75360714 [T/C]	0.06 [0.04, 0.07]	1.18×10^{-12}	chr5:74656539 [T/C]	0.06 [0.06, 0.06]	5.46×10^{-173}
chr6:160589086 [A/G]	0.12 [0.09, 0.16]	2.94×10^{-12}	chr6:161010118 [A/G]	0.08 [0.07, 0.09]	4.80×10^{-79}
chr19:11085680 [AC/A]	-0.13 [-0.15, -0.10]	3.73×10^{-27}	Not in prior GWAS		
chr19:44908684 [T/C]	0.18 [0.16, 0.20]	9.51×10^{-62}	chr19:45411941 [T/C]	0.18 [0.18, 0.19]	9.12×10^{-828}
chr19:44908822 [C/T]	-0.40 [-0.42, -0.37]	2.03×10^{-181}	chr19:45412079 [C/T]	-0.48 [-0.48, -0.47]	2.91×10^{-3040}
chr19:44935906 [C/G]	-0.03 [-0.05, -0.02]	1.46×10^{-4}	chr19:45439163 [C/G]	-0.02 [-0.03, -0.01]	1.96×10^{-10}

The summary statistics in prior GWAS were extracted from the GWAS used in genetic correlation analysis. When the SNV was not available in the reference GWAS, the summary statistics were obtained from the GWAS catalog. SNVs were noted as not available if it has not been reported in the GWAS catalog. Effect size: odds ratios (OR) for disease phenotypes, beta for continuous traits.

Supplementary Methods

Study population

One of the goals set by the *All of Us* research program was to recruit individuals that have been and continue to be underrepresented in biomedical research due to limited access to health care. Therefore, *All of Us* takes demographics, including race, ethnic group, age, sex, gender identity, income, educational attainment, and geographic location, into account when enrolling participants.¹ For the first release of the genomic data, *All of Us* prioritized historically underrepresented individuals in the sequencing procedure, resulting in a 51% percentage of participants in racial and ethnic minorities among the 98,622 sequenced individuals. The detailed ancestry summary statistics (both genetically predicted and self-reported) are presented in **Table 1**. All participants completed electronic consent modules and health questionnaires upon enrollment, and the study protocol has been published previously.² In the current release, all samples with genetic data have at least one other type of data that can be used for research purposes. Approval to use the dataset for the specified demonstration projects was obtained from the *All of Us* Institutional Review Board.

Cardiometabolic phenotypes

The cardiometabolic traits included in the present study were atrial fibrillation (AF), coronary artery disease (CAD), type 2 diabetes (T2D), body height, and low-density lipoprotein (LDL). AF was defined using a combination of International Classification of Diseases (ICD) codes, self-reported personal medical history, and procedure and operation codes. The detailed

algorithm for AF is described in **Table S2**. For CAD and T2D, we used published phenotype algorithms obtained from the eMERGE network to define the disease phenotypes using electronic health records (EHR) data.^{3,4} The CAD algorithm was based on ICD and CPT codes, and the T2D algorithm used information from ICD codes, medication use, and laboratory test results. Body height was extracted from the program's physical measurements data section, which includes data for blood pressure, height, weight, waist circumference, hip circumference, and heart rate, all were measured at enrollment by the program. LDL cholesterol levels were extracted from EHR with the unit of mg/dL. When there were multiple laboratory results for LDL in a participant's EHR, the most recent record was used. Rank-based inverse normal transformation was applied to continuous traits before association testing.

Whole-Genome Sequencing and variant calling

Each Genome Center performed quality control (QC) of the specimens obtained from the *All of Us* Biobank. Sample preparation and normalization and DNA library construction have been reported previously.⁵ The Illumina NovaSeq 6000 instrument was used to conduct the whole-genome sequencing (WGS) procedure following the manufacturer's best practices. Post-sequencing analysis was performed using Illumina's DRAGEN pipeline, which was harmonized (v3.4.12) between different Genome Centers. The GRCh38 reference genome was used in the alignment step. The single sample QC processes checked fingerprint concordance (array vs. WGS data), sex concordance (genetically determined vs. self-reported), cross-individual contamination rate and coverage to detect major errors, such as sample swaps or contamination. Participants who failed these tests were removed from the current release. The WGS variants were called jointly to reduce systematic biases. Additional sample QC procedures were then performed on the joint callsets, including hard threshold flagging (e.g., number of SNPs: < 2.4M

and > 5.0M) and population outlier flagging. Variants QC was performed after sample QC, flagging specific variants from a callset. Processes included hard threshold filters (e.g., ExcessHet, QUAL score) and Allele-Specific VariantQualityScoreRecalibration (AS-VQSR or VQSR, a machine learning technique for identifying variants that are likely artifacts).

Variant-level and sample-level quality control

In addition to the QC procedures performed by the program when producing the genomic data, we applied several variant-level and sample-level filterings to keep only high-quality data in the present analysis. For genetic variants that passed internal QC, we further filtered out monomorphic variants, variants in low-complexity regions, variants with call rate < 95%, and variants with Hardy-Weinberg equilibrium P value < 1×10^{-15} . We conducted sample QC by excluding samples with call rate < 95%, Ti/Tv ratio, het/hom ratio, SNP/Indel ratio, or number of singletons > 8 standard deviation (SD) from the population mean. We performed the QC procedures using PLINK 2.0 [<https://www.coggenomics.org/plink/2.0/>].

Principal components

Population structure inference on the entire study population was obtained by implementing an algorithm (PC-AiR) that accounts for relatedness in the sample.⁶ This method takes the kinship inference obtained using the KING software⁷ as input, which assigns negative estimates to pairs of individuals with different ancestry backgrounds. PC-AiR uses these negative kinship estimates to identify groups of participants with different ancestry backgrounds and performs principal component analysis (PCA) on unrelated samples who are representative of the ancestries presented in the entire sample. For the related subset, the algorithm predicts PC values for them based on their genetic similarities with the unrelated subset. Genetic similarities were also

estimated using the kinship inference algorithm that accounts for unknown population substructure.⁷ We followed a 2-step method presented in the TOPMed analysis pipeline (https://uw-gac.github.io/topmed_workshop_2017/index.html) to calculate the principal components. R package *GENESIS* was used to implement the PC-AiR and PC-Relate algorithms needed for this analysis.

Common variant analysis and genetic correlation

We performed association testing for phenotypes and individual genetic variants with minor allele frequencies (MAF) > 1% using a whole-genome regression approach implemented in the REGENIE⁸ software assuming an additive genetic model. This method first generates LOCO (leave-one-chromosome-out) predictions of trait values in step 1 using a set of high-quality genetic variants that provide whole-genome information. We used genetic variants with a MAF > 1%, minor allele count (MAC) > 100, missingness rate < 1%, Hardy-Weinberg equilibrium test P value > 1×10^{-15} and linkage-disequilibrium (LD) pruning ($r^2 < 0.1$). The resulting LOCO predictions are then used in step 2 to test the associations between phenotypes and each individual genetic variant. The covariates adjusted in all statistical models were: (1) age (age at enrollment for disease phenotypes, age at measurement for continuous traits), (2) sex, (3) 20 principal components of ancestry. Age at enrollment was calculated using birthdate and program consent date. When consent date was missing from the database, we used median consent date among the samples to calculate the enrollment age. 57,239 participants have consent date information available in the database, ranging from 2017-05-31 to 2021-04-01 (median date: 2019-03-08).

We fitted linear regression models for continuous traits and logistic regression models for binary traits with a saddle point approximation (SPA) method⁹ accounting for case-control

imbalance. Genome-wide significant variants were considered those with $P < 5 \times 10^{-8}$. We also conducted conditional analyses¹⁰ using the GCTA software to select secondary independent significant SNVs at each locus. The mapped gene at each locus was either the nearest gene or the phenotype-associated gene within 500 kb range of the lead variant reported in the GWAS catalog. We then estimated genetic correlation using GWAS summary statistics to evaluate how consistent our GWAS results were with corresponding previously published GWAS results. We used the LD score regression approach¹¹ implemented in the LDSC software and pre-computed LD-scores provided by the software for this analysis. LDSC is not a bounded estimator and thus may generate estimates less than -1 or greater than 1 due to sampling variation. Also, a genetic correlation estimate that is greater than 1 may indicate that the true genetic correlation is high and there are some samples that overlap between the two studies. All correlation estimates for disease phenotypes were converted from observed scale to liability scale using sample prevalence and population prevalence estimates.

Rare variant analysis using burden test

To test the association between each phenotype and the burden of rare variants (MAF < 0.1%) of their known associated genes, we first annotated the protein consequences of rare variants using the Loss-of-Function Transcript Effect Estimator (LOFTEE)¹² plug-in implemented in the Variant Effect Predictor¹⁰⁵ (VEP; v.95) (<https://github.com/konradjk/loftee>) software to identify high-confidence loss-of-function (LoF) variants, including frameshift indels, stop-gain variants and splice site disrupting variants. Furthermore, we checked the continental allele frequencies using gnomAD v2 to make sure the maximum population frequency (POP_MAX) for each of the rare variations was also < 0.1% since ancestral allele frequencies can be different from the pooled ones. We identified 597 LoF variants within *TTN*, 38 within *GIGYF1*, 32 within *APOB*,

17 within *LDLR*, 14 within *PCSK9*, 49 within *ADAMTS17*, 15 within *ACAN*, and 32 within *NPR2*. We then tested the association between the burden of selected rare variants in each gene and its corresponding reported phenotype using a burden test implemented in the REGENIE software.⁸ Specifically, for each gene, we counted the number of alternative allele copies each individual carries and treated this number as a single burden genotype. This genotype was then used in a regression model (linear models for continuous traits, logistic models for binary traits) to associate with the phenotype, adjusting for age (age at enrollment for disease phenotypes, age at measurement for continuous traits), sex, and 20 principal components of ancestry. The LOCO predictions obtained from common variant analysis were used in step 2 for this analysis. We also applied the saddle point approximation (SPA) method⁹ to account for case-control imbalance.

References

1. The “All of Us” Research Program | NEJM. Accessed November 1, 2021. <https://www.nejm.org/doi/full/10.1056/NEJMSr1809937>
2. The All of Us Research Program Initial Protocol. National Institutes of Health (NIH) — All of Us. Published August 11, 2020. Accessed March 19, 2022. <https://allofus.nih.gov/news-events/announcements/all-us-research-program-initial-protocol>
3. MidSouth CDRN - Coronary Heart Disease Algorithm | PheKB. Accessed November 2, 2021. <https://phekb.org/phenotype/midsouth-cdrn-coronary-heart-disease-algorithm>
4. Type 2 Diabetes Mellitus | PheKB. Accessed November 2, 2021. <https://www.phekb.org/phenotype/type-2-diabetes-mellitus>
5. Venner E, Muzny D, Smith J, et al. *Whole Genome Sequencing as an Investigational Device for Return of Hereditary Disease Risk and Pharmacogenomic Results as Part of the All of Us Research Program*. *Genetic and Genomic Medicine*; 2021. doi:10.1101/2021.04.18.21255364
6. Conomos MP, Miller MB, Thornton TA. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet Epidemiol*. 2015;39(4):276-293. doi:10.1002/gepi.21896
7. Robust relationship inference in genome-wide association studies | Bioinformatics | Oxford Academic. Accessed November 2, 2021. <https://academic.oup.com/bioinformatics/article/26/22/2867/228512>
8. Mbatchou J, Barnard L, Backman J, et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet*. 2021;53(7):1097-1103. doi:10.1038/s41588-021-00870-7
9. Zhou W, Nielsen JB, Fritsche LG, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet*. 2018;50(9):1335-1341. doi:10.1038/s41588-018-0184-y
10. Yang J, Ferreira T, Morris AP, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet*. 2012;44(4):369-375. doi:10.1038/ng.2213
11. Bulik-Sullivan BK, Loh PR, Finucane HK, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet*. 2015;47(3):291-295. doi:10.1038/ng.3211
12. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17(1):122. doi:10.1186/s13059-016-0974-4