**Supplementary Materials**

# Jurgens, et al.

## Description of Participating Studies

The **UK Biobank** (UKB) is a large population-based prospective cohort study from the United Kingdom that included over 500,000 individuals with deep phenotypic data, including medical interviews, electronic health record linkage and death registry linkage[1,2]. Participants were recruited between 2006 and 2010 at ages of 40-69 years[2]. Relevant genomic data currently includes exome sequencing on over 470,000 samples[3] and genome sequence data on approximately 490,000 samples[4]. AF was defined using nurse-assisted self-report, in-patient diagnosis codes, and procedural codes, as described previously[5]. The UK Biobank resource was approved by the UK Biobank Research Ethics Committee and all participants provided written informed consent to participate. Use of UK Biobank data was performed under application number 17488 and was approved by the Mass General Brigham Institutional Review Board.

The **Geisinger MyCode** study, also known as the DiscovEHR study[6,7]. Started in 2007, the study is open to any Geisinger patient—through opt-in informed consent—including both primary and specialty care clinics, and has enrolled over 360,000 participants to date. Through the DiscovEHR collaboration with Regeneron Genetics Center, whole-exome sequencing from collected blood samples has been completed for approximately 175,000 participants to date, and linked with health information from the Geisinger electronic health record (1996–present). This study leveraged exome data for 153,338 adult (≥18) individuals uniformly sequenced using an IDT exome capture platform and who passed subsequent central quality control procedures. AF status was defined using a combination of 15 ICD10-CM, ICD9-CM and CPT4 codes (I48, I48.0, I48.1, I48.2, I48.3, I48.4, I48.9, I48.91, I48.92, 427.3, 427.31, 427.32, 93656, 93650, 93657) aggregated from Geisinger's Electronic Health Records (EHR). The Geisinger Institutional Review Board approved the MyCode project and the present analysis.

The NIH's **All of Us Research Program** is a longitudinal cohort study that aims to include 1 million racially, ancestrally and demographically diverse participants from across the United States, combining phenotypic data from various sources including patient-derived information and electronic health record linkage[8]. One of the goals was to recruit individuals that have been and continue to be underrepresented in biomedical research because of limited access to health care[9]. Consistently, All of Us prioritized underrepresented participants for genome sequencing and data collection and

included them in the first few releases of the dataset, resulting in a diverse research population with rich phenotypic data. As part of the V7 release in April 2023, whole genome sequencing (WGS) was performed on approximately 250,000 participants using Illumina NovaSeq 6000 machines following manufacturer's best practices. Same protocol for library preparation (PCR Free Kapa HyperPrep) and software for variant calling (DRAGEN v3.4.12) were used to keep consistent WGS data generated from different Genome Centers. A stringent central QC procedure was applied, as described in the program's genomic quality report [https://support.researchallofus.org/hc/en-us/articles/4617899955092-All-of-Us-Genomic-Quality-Report-], leaving 245,394 samples (47.7% described as racial/ethnic minorities). The V7 data release was used for all main burden testing discovery analyses, as described in the **Supplementary Note** sections below. AF patients were identified using the centrally curated definition, which included SNOMED and ICD codes mapping to the following OMOP concepts: 313217, 4154290, 605092, 4141360, 4232697, 45768480, 4232691, 4108832, 44782442, 605092, 4117112, 4119601, 4199501, 314665, 36714994, 36712986, 4108832, 4146580, 4137382.

In Februari 2024, the V8 dataset of AllofUs was released, including data on 414,830 individuals after internal QC procedures (https://support.researchallofus.org/hc/en-us/articles/29390274413716-All-of-Us-Genomic-Quality-Report). The V8 dataset was used for downstream analyses pertaining to heritability and cross-ancestry effect sizes. Furthermore, non-overlapping samples were used for replication analyses of burden results, as described in more detail in the **Supplementary Note** sections below.

All enrolled participants provided informed consent to AllofUs. Use of AllofUs data was approved under a data use agreement between the Massachusetts General Hospital and the All of Us research program.

The **Atherosclerosis Risk in Communities (ARIC)** study is a prospective population-based study of 15,792 men and women 45 to 64 years of age at enrollment (73% of European descent), recruited from four communities in the United States (suburbs of Minneapolis, Minnesota; Washington County, Maryland; Jackson, Mississippi; and Forsyth County, North Carolina) between 1987-1989 to investigate the epidemiology of cardiovascular disease. Participants underwent electrocardiograms at baseline and at each follow-up exam (3 exams; 1 exam every 3 years). Incident atrial fibrillation was classified as the first occurrence of atrial fibrillation as identified from electrocardiograms at study

visits, hospital discharge codes (ICD-9-CM code 427.31 or 427.32) or death certificates (ICD-9 code 427.3 or ICD-10 code I48).

The **Mount Sinai BioMe Biobank (BioMe)** is an ongoing, prospective, hospital- and outpatient-based population research program operated by The Charles Bronfman Institute for Personalized Medicine (IPM) at Mount Sinai. BioMe has enrolled over 50,000 participants between September 2007 and July 2019. BioMe is an Electronic Medical Record (EMR)-linked biobank that integrates research data and clinical care information for consented patients at The Mount Sinai Medical Center, which serves diverse local communities of upper Manhattan with broad health disparities. IPM BioMe populations include 25% of African American ancestry (AA), 36% of Hispanic Latino ancestry (HL), 30% of white European ancestry (EA), and 9% of other ancestry. The BioMe disease burden is reflective of health disparities in the local communities. BioMe operations are fully integrated in clinical care processes, including direct recruitment from clinical sites waiting areas and phlebotomy stations by dedicated BioMe recruiters independent of clinical care providers, prior to or following a clinician standard of care visit. Recruitment currently occurs at a broad spectrum of over 30 clinical care sites.

The **BioVU** is Vanderbilt's biobank of DNA extracted from leftover and otherwise discarded clinical blood specimens. BioVU operates as a consented biorepository; all individuals must sign the BioVU consent form in order to donate future specimens. BioVU subjects are de-identified and linked to the Synthetic Derivative enabling researchers to access genetic data/DNA material as well as dense, longitudinal electronic medical record (EMR) information.

**Cleveland Clinic Lone Atrial Fibrillation GeneBank Study (CCAF)** has enrolled patients with lone atrial fibrillation, defined as atrial fibrillation in the absence of significant structural heart disease. Participants were at least 18 years of age with a history of recurring or persistent lone atrial fibrillation, ≤ 50% coronary artery stenosis in the coronary arteries (if cardiac catheterization done) or with normal stress test results (documentation of normal cardiac catheterization or stress test required if age ≥ 50 years), and had normal left ventricular ejection fraction (LVEF) 50%. Individuals were excluded if they had heart failure, history of significant valvular disease (>2+ valvular regurgitation, any valvular stenosis), significant coronary artery disease (>50% coronary artery stenosis), prior myocardial infarction, prior percutaneous coronary intervention, or coronary artery bypass graft, or latest LVEF <50%.

**The Cleveland Family Study (CFS)** was designed to investigate the familial basis of sleep disordered breathing. The study consists of data collected from 2,284 individuals (46% African-American) from 361 families. Index probands (n=275) were recruited from 3 area hospital sleep labs if they had a confirmed diagnosis of sleep apnea and at least 2 first-degree relatives available to be studied. In the first 5 years of the study, neighborhood control probands (n=87) with at least 2 living relatives available for study were selected at random from a list provided by the index family and also studied. All available first-degree relatives and spouses of the case and control probands also were recruited. Second-degree relatives, including half-sibs, aunts, uncles and grandparents, were also included if they lived near the first-degree relatives (cases or controls), or if the family had been found to have two or more relatives with sleep apnea. Blood was sampled and DNA isolated for participants seen in the last two exam cycles. The sample, which is enriched with individuals with sleep apnea, also contains a high prevalence of individuals with sleep apnea-related traits, including obesity, impaired glucose tolerance, and hypertension. Four visits occurred from 1990 - 2006, including a final visit from 2001 - 2006 at a General Clinical Research Center. The last three exams targeted all subjects who had been studied at earlier exams, as well as new minority families and family members of previously studied probands who had been unavailable at prior exams.

**The Cardiovascular Health Study (CHS)** is a population-based cohort study of risk factors for coronary heart disease and stroke in adults ≥65 years conducted across four field centers (Sacramento, CA; Hagerstown, MD; Winston-Salem, NC; Pittsburgh, PA). The original predominantly European ancestry cohort of 5,201 persons was recruited in 1989-1990 from random samples of the Medicare eligibility lists; subsequently, an additional predominantly African-American cohort of 687 persons was enrolled for a total sample of 5,888. Blood samples were drawn from participants during annual clinic visits. DNA was subsequently extracted from available samples. CHS participants selected for inclusion in the TOPMed sequencing program included African-American participants, cases of idiopathic venous thromboembolism, myocardial infarction, coronary heart disease, or stroke along with a random sample of “healthy elderly”. CHS was approved by institutional review committees at each field center and individuals in the present analysis had available DNA and gave informed consent including consent to use of genetic information for the study of cardiovascular disease.

**COPDGene** is a multicenter observational study designed to identify genetic and clinical factors associated with COPD. COPDGene has enrolled more than 10,000 current and former heavy smokers across all stages of pulmonary functions. Both Non-Hispanic white and African-American subjects are included in the cohort. Inspiratory and expiratory chest CT scans have been obtained on all participants. In addition to the cross-sectional enrollment process, these subjects are being followed regularly with 5 year complete study visits as well as semi-annual questionnaires for longitudinal studies.

In the **Australian Familial AF** Study, a cohort of probands with familial AF was recruited for genetics studies at the Victor Chang Cardiac Research Institute. Familial AF cases were identified from in-patient and out-patient populations at St. Vincent's Hospital and by referral from collaborating physicians throughout Australia. Study subjects underwent clinical evaluation with history, ECG and echocardiogram, and informed consent was obtained from all participants. 151 probands aged <66 years at the time of diagnosis were included in this analysis. The control cohort was comprised of age- and sex-matched individuals (n=151) who had no history of cardiovascular disease.

The **CATHeterization GENetics (CATHGEN)** biorepository collected biospecimens and clinical data on individuals age ≥18 undergoing cardiac catheterization at a single center (Duke University Medical Center) from 2000-2010; a total of N=9334 individuals were collected. Samples were matched at the individual level to clinical data collected at the time of catheterization and stored in the Duke Databank for Cardiovascular Diseases (DDCD). Clinical data included subject demographics, cardiometabolic risk factors, cardiac history including symptoms, age-of-onset of cardiovascular diseases, coronary anatomy and cardiac function at catheterization, laboratory data, and yearly follow-up for hospitalizations, vital status, medication use and lifestyle factors. AF cases were defined as individuals who had ever had AF based on any ECG available at Duke University or ICD-9 code for AF used for inpatient or outpatient billing.

The **Framingham Heart Study (FHS)** is a community-based observational cohort initiated in 1948 to prospectively investigate CVD and its risk factors. The Original cohort (n=5,209) received biennial exams. The Original Cohort children (& spouses), termed the Offspring cohort (n=5,214), were recruited in 1971, and have been examined every four to eight years. In FHS, all cardiovascular hospital and outside records were routinely obtained and electrocardiograms were recorded at all FHS examinations; atrial fibrillation cases through 2017 were verified by two FHS cardiologists.

The **Groningen Genetics of Atrial Fibrillation (GGAF)** cohort is a cohort composed from 5 different sources of individuals with atrial fibrillation (AF) and age and sex-matched controls. Written informed consent was provided from all participating individuals, and all 5 studies were approved by the ethical committee at the University Medical Center Groningen and Maastricht University Medical Center. AF was documented by ECG. All participants underwent clinical examination, including ECG, echocardiography (only AF cases), and blood draw, from which DNA has been extracted.

The **RACE 3 tria**l is a randomized, prospective multicentre, open-label, blinded endpoint trial (Clinicaltrials.gov identifier NCT00877643). Patients with a short history of symptomatic persistent AF [total AF history <5 years, total persistent AF duration >7 days but <6months, ≤1 electrical cardioversion (ECV)] and mild to moderate stable HF were included. The study has been performed in compliance of the Declaration of Helsinki. The study was approved by The Institutional Review Board of all participating hospitals, and all patients gave written informed consent. Patients were randomized to targeted therapy of underlying conditions or conventional therapy. All patients received treatment according to the AF and HF guidelines and were treated with rhythm control therapy. Patients were scheduled for ECV 3 weeks after inclusion. If AF relapsed, repeat ECV, antiarrhythmic drugs, and atrial ablations were allowed. In addition, the targeted therapy group received four therapies: (i) angiotensin-converting enzyme inhibitors (ACE-Is) and/or angiotensin receptor blocker (ARBs), (ii) statins, (iii) mineralocorticoid receptor antagonists (MRAs), and (iv) cardiac rehabilitation.

The **RACE V** trial is an ongoing investigator-initiated, prospective, multicentre registry aiming to include 750 patients in multiple centres in The Netherlands. Inclusion criteria were patients aged ≥18 years with paroxysmal AF, a maximum AF history of 10 years since diagnosis at the moment of inclusion, a maximum CHA2DS2-VASc score of 5, and no other indication for anticoagulation drugs (e.g. mechanical valve prosthesis). Patients had to have at least two documented episodes of paroxysmal AF in the past year or one documented episode with at least two symptomatic episodes in the past year suspected to be AF without documentation. In patients with a Medtronic pacemaker, atrial high rate episodes (AHREs) >190 beats per minute lasting >6 min were qualified as AF episodes. Patients with other types of pacemakers, defibrillators or cardiac resynchronization therapy could not participate due to differences in AHRE algorithm and/or incompatibility with the type of home-monitoring. Further exclusion criteria were patients with a history of persistent AF, currently on amiodarone, current pregnancy or a life expectancy <2.5 years, patients with AF caused exclusively

due to transient triggers (e.g. postoperative, due to infection), patients with a previous pulmonary vein isolation (PVI), or intention to undergo PVI, or diagnosed congenital heart disease. In total, 247 patients had available blood samples; from this amount, a total of 34 (14%) were excluded because of AF at the moment of sampling. Samples from the remaining 213 patients were used for the current analyses. RACE V was performed in concordance with the Declaration of Helsinki. The Institutional Review Board approved the protocol.

The **BioHEART study** is a longitudinal, prospective cohort study, aiming to recruit more than 10,000 adult participants undergoing clinically referred CTCA. After informed consent, participant data, blood samples and CTCA imaging data are recorded. Follow-up for all patients is conducted 1 month after recruitment, and then annually for the life of the study. CTCA data provide volumetric quantification of total calcified and non-calcified plaque, which will be assessed using established and novel scoring systems. Comprehensive molecular phenotyping is performed using state-of-the-art genomics, transcriptomics, metabolomics, proteomics, lipidomics, and immunophenotyping.

The **University of Illinois Chicago AF registry** is an ongoing registry of AF patients enrolled through the University of Illinois Chicago, as a continuous prospective study focusing on clinical outcomes and genetics in adults with AF. Patients with a history of AF associated with cardiothoracic surgery were excluded. We obtained information on demographics, cardiovascular risk factors, and family history for baseline characteristics with blood drawn for DNA extraction at the time of enrollment.

The **Heart and Vascular Health Study (HVH)** is a case-control study of risk factors for development of cardiovascular disease, conducted in the setting of Kaiser Permanente Washington (formerly Group Health Cooperative), a large integrated healthcare system in Washington State, USA. In the AF study, plan members assigned a new ICD-9 code of 427.31 or 427.32 in the inpatient or outpatient setting during 2001-2007 were identified. Incident AF was verified by review of medical records with the requirement that the AF be documented by 12-lead electrocardiogram and clinically recognized by a physician, with no previous evidence of AF in the medical record. The date that AF was clinically recognized was recorded as the index date for each AF case. This analysis included only AF cases 30 to less than 61 years of age at the index date who self-identified as white; no control subjects from the HVH study were included.

The **Jackson Heart Study (JHS)** recruited 5,306 African American participants from the Jackson, Mississippi, metropolitan tri-county area (Hinds, Madison, and Rankin) between 2000 and 2004. JHS is a prospective, community-based cohort designed to investigate risk factors for cardiovascular disease among African Americans. A range of measures, including traditional and putative CVD risk factors, health behaviors, detailed demographic, socioeconomic and sociocultural factors, medication use, anthropometry, blood pressure, assessments of kidney function and diabetes, and biochemical analytes, were obtained at the baseline JHS examination and in two subsequent clinic visits (2005-2008 and 2009-2013), with a fourth exam also recently completed. Biological samples (i.e., blood and urine) have been assayed for putative biochemical risk factors and stored for future research. DNA has been extracted and lymphocytes have been cryopreserved for studies of candidate genes, genome-wide scanning, whole genome sequencing, expression, and other –omics investigations. Atrial fibrillation status was defined based on ECG's at JHS visits 1 and 3, based on Minnesota code 8-3-1. Age at diagnosis is set to age at visit.

The **Johns Hopkins University School of Medicine Atrial Fibrillation Genetics Cohort (JHU_AF)** started collecting data in September 2008 on patients >18 years of age that were referred for catheter ablation of symptomatic drug refractory atrial fibrillation. De-identifiable patient samples and clinical information were collected prior to ablation procedures.

The **Massachusetts General Hospital AF (MGH_AF)** study has enrolled serial patients with early-onset atrial fibrillation referred to the Arrhythmia Service from July 5, 2001 onwards. Atrial fibrillation was documented by electrocardiography. Individuals with structural heart disease as assessed by echocardiography, hyperthyroidism, myocardial infarction, or heart failure were excluded. Each patient underwent a physical examination and standardized interview. All patients were evaluated by 12-lead electrocardiogram, echocardiogram, and laboratory studies.

The **Malmö Preventive Project (MPP)** was a community-based disease prevention program including 33,346 inhabitants from the city of Malmö in Southern Sweden. Complete birth cohorts between 1921-1949 were invited, and the participation rate was 71%. Participants underwent screening between 1974 to 1992 for cardiovascular risk factors, alcohol abuse, and breast cancer. Between 2002-2006, surviving participants were invited to a reexamination which included blood sampling from which DNA has been extracted. Subjects with prevalent or incident AF were identified from national registers as previously described, and cases with DNA were then matched in a 1:1

fashion to controls with DNA from the same cohort by sex, age (±1 year), and date of baseline exam (±1 year). Also, controls required a follow-up exceeding that of the corresponding AF case.

The **Atrial Fibrillation Biobank Ludwig Maximilian University (AFLMU) Study** contributes to the spectrum of disease by adding carefully characterized patients with atrial fibrillation. Atrial fibrillation, one of the most common human arrhythmias confers major morbidity, mortality and health care cost, and has been demonstrated to be caused and influenced by genetic and -omics factors. Particularly, AFLMU enrolled patients with an early onset of atrial fibrillation to increase the genetic burden on disease pathophysiology. All patients were recruited applying standardized protocols to maintain homogeneity in data and DNA quality.

The **Mass General Brigham Biobank** (MGB; previously Partners HealthCare Biobank) is a biorepository of consented patient samples at Partners HealthCare (parent organization of Massachusetts General Hospital and Brigham and Women's Hospital). Aliquots of DNA, plasma, and serum are linked to electronic medical record data and are available to investigators for use.

For TOPMed, select cases with early-onset AF from MGB were identified using a validated electronic medical record search algorithm that has been previously described. Briefly, the algorithm is based on diagnostic codes, procedure codes, electrocardiographic information, and medications. The charts of individuals with an electronic medical record indicator of white race, and with an electronically ascertained AF onset <61 years of age, were manually reviewed by a research nurse for validation. Those with cardiac surgery, cardiomyopathy, myocardial infarction, or valvular heart disease prior to AF onset were excluded from consideration for sequencing in the TOPMed Program.

A subset of ~53k MGB participants have undergone exome sequencing in a separate effort. An initial freeze of ~30k MGB participants were included in the CCDG-WES call set (see below), while the remaining participants were included in a separate call set.

The **Penn Biobank atrial fibrillation (PMBB_AF)** study selected the cases with early-onset atrial fibrillation were selected from the Penn Biobank based on an age of atrial fibrillation onset prior to 61 years of age, and in the absence of a myocardial infarction, heart failure or severe valvular disease.

The **DECAF** trial was conducted at the Texas Cardiac Arrhythmia Institute (TCAI) in 2013 in collaboration with the University of Texas at Austin. Four hundred consecutive AF patients undergoing catheter ablation were enrolled. All participants provided voluntary informed consents. Blood samples were collected before the ablation procedure and labeled with anonymous patient identifier. The researchers at UT Austin responsible for DNA extraction and genetic analysis were blinded about the clinical characteristics and identification of the study participants. AF cases included adults >18 years of age from both sex and all AF types.

The **University of California San Francisco atrial fibrillation study (UCSF_AF)** built on Cardiovascular Research Institute (CVRI) Resource in Arteriosclerosis and Metabolic Disease is an ongoing multi-ethnic study of adults ≥ 18 years of age which was started in 1989 and now includes 28,000 participants recruited from the UCSF medical system. Within the Resource lies data and biospecimens from nearly 1,000 patients presenting to the electrophysiology laboratory for electrophysiology procedures that were densely phenotyped for electrophysiologic characteristics with biospecimens collected from various intra and extra-cardiac chambers. Phenotyping of all participants was achieved via interview and review of medical records.

The University of Massachusetts Medical School of **miRhythm** Study is an ongoing study of adult patients undergoing an elective electrophysiology study or arrhythmia ablation procedure for a supraventricular or ventricular arrhythmia, including atrial fibrillation (AF). Atrial fibrillation is a major clinical and public health problem that is related to atrial pathologic remodeling. Few tools are available to quantify the activity or extent of this remodeling, rendering it difficult to identify individuals at risk for AF. Previous studies have suggested an important role for miRNA in cardiovascular disease through gene expression regulation, making this a promising avenue for studying AF mechanisms.

The **Vanderbilt Atrial Fibrillation Registry (VAFR)** was founded in 2001. Patients with AF and family members are prospectively enrolled. At enrollment, a detailed past medical history is obtained along with an AF symptom severity assessment. Blood samples are obtained for DNA extraction. Patients are followed longitudinally along with serial collection of AF symptom severity assessments.

The **Vanderbilt Atrial Fibrillation Ablation Registry (VAFAR)** is a prospective observational registry of subjects undergoing AF ablation (clinicaltrials.gov NCT #02404415). Written informed consent is obtained prior to ablation. DNA is extracted from whole blood collected during the procedure.

Baseline clinical data is manually extracted from the medical record and supplemented by patient interview. Participants are prospectively followed for arrhythmia recurrence post-ablation according to current guidelines.

The **Women's Genome Health Study (WGHS)** is a prospective cohort comprised of over 25,000 initially healthy female health professionals enrolled in the Women's Health Study, which began in 1992-1994. All participants in WGHS provided baseline blood samples and extensive survey data. Women who reported atrial fibrillation during the course of the study were asked to report diagnoses of AF at baseline, 48 months, and then annually thereafter. Participants enrolled in the continued observational follow-up who reported an incident AF event on at least one yearly questionnaire were sent an additional questionnaire to confirm the episode and to collect additional information. They were also asked for permission to review their medical records, particularly available ECGs, rhythm strips, 24-hour ECGs, and information on cardiac structure and function. For all deceased participants who reported AF during the trial and extended follow-up period, family members were contacted to obtain consent and additional relevant information. An end-point committee of physicians reviewed medical records for reported events according to predefined criteria. Incident AF was confirmed if there was ECG evidence of AF or if a medical report clearly indicated a personal history of AF. The earliest date in the medical records when documentation was believed to have occurred was set as the date of onset of AF.

The **Genetics in AF (GENAF)** study enrolled individuals with early-onset lone AF before age 50 in Norway between 2009 and 2016. Early-onset was defined as diagnosis of AF before age 50. Lone AF was defined as AF in the absence of clinical or echocardiographic findings of cardiovascular disease, hypertension, metabolic or pulmonary disease. AF was documented in ECG. All participants underwent clinical examination, including ECG, echocardiography, and blood draw, from which DNA has been extracted. The study conforms to the principles of the Declaration of Helsinki and was approved by the Regional Ethics Committee (REK) in Norway (Protocol reference number: 2009/2224-5). All included patients gave written informed consent.

**The Intermountain Healthcare Biological Samples Collection Project and Investigational Registry for the On-going Study of Disease Origin, Progression and Treatment (INSPIRE_AF)'s** purpose is to collect biological samples, clinical information and laboratory data from Intermountain Healthcare patients. The registry originally collected samples in patients undergoing coronary

angiography as part of the Intermountain Heart Collaborative Study. It has been expanded to collect samples in patients diagnosed with all types of medical conditions and patients from the general population including those who have not been diagnosed with health related issues. Just over 25,000 individuals have provided samples as part of this registry. The registry enables researchers to develop a comprehensive collection of information that may help in disease management, including determining best medical practices for predicting, preventing and treating medical conditions.

**The early Treatment of Atrial Fibrillation for Stroke Prevention Trial (EAST-AFNET 4)** was designed to test whether a strategy of early rhythm-control therapy that includes atrial fibrillation ablation would be associated with better outcomes in patients with early atrial fibrillation than contemporary, evidence-based usual care.

**ENGAGE AF-TIMI 48** was a multinational, randomized, double-blind, double-dummy trial comparing 2 dosing regimens of edoxaban with warfarin[10,11]. The trial was conducted at 1393 centers in 46 countries. Patients were enrolled during the period from November 19, 2008, through November 22, 2010. Eligible patients were ≥21 years of age with AF within 12 months and a CHADS2 risk score ≥2. Key exclusion criteria included a CrCl <30 mL/min estimated with the Cockcroft-Gault method, a high risk of bleeding, and the use of dual antiplatelet therapy.The ethics committee at each participating center approved the protocol. Written informed consent was obtained from all patients. Patients (n=21 105) were randomly assigned in a 1:1:1 ratio to receive warfarin, a high-dose edoxaban regimen (HDER; 60 mg daily or 30 mg daily if a dose reduction was required), or a lower-dose edoxaban regimen (LDER; 30 mg or 15 mg daily if a dose reduction was required).

**SAVOR-TIMI 53** was a multicenter, randomized, double-blind, placebo-controlled trial that randomized 16 492 patients at 788 sites in 26 countries from May 2010 to December 2011 with type 2 DM, hemoglobin A1c between 6.5% and <12.0% within 6 months of randomization, and either a history of established cardiovascular disease or multiple risk factors for vascular disease to receive either 5 mg of saxagliptin daily (or 2.5 mg daily in patients with an estimated glomerular filtration rate [eGFR] of ≤50 mL/min) or matching placebo. The full eligibility criteria and analysis plan have been reported previously[12,13]. Written informed consent was obtained from all patients. The relevant ethics committees at all participating centers approved the protocol.

**PEGASUS-TIMI 54** trial, 21 162 patients - from 1161 sites in 31 countries - with a history of a spontaneous myocardial infarction and 1 additional high-risk feature were randomly assigned to receive 90 mg of ticagrelor orally twice daily, 60 mg of ticagrelor orally twice daily, or placebo for the duration of the trial. Full inclusion and exclusion criteria have been published previously[14,15]. Stable patients were recruited a median of 1.7 years (interquartile range, 1.2-2.3 years) from their qualifying myocardial infarction. All patients were to receive aspirin at a low dose of 75 to 150 mg daily. Patients with known hemorrhagic diathesis or a coagulation disorder, a history of an ischemic stroke or previous intracranial bleeding at any time, gastrointestinal bleeding within the past 6 months, or major surgery within 30 days were excluded. This study was approved by the corresponding health authorities and ethics board or institutional review boards for all participating study sites, and the patients provided written informed consent to participate in the trial.

**FOURIER (TIMI 59)** was a randomized, double-blind, placebo-controlled clinical trial that enrolled 27 564 patients from 1242 sites in 49 countries, aged 40 to 85 years with clinically evident atherosclerotic cardiovascular disease (prior myocardial infarction, prior nonhemorrhagic stroke, or symptomatic peripheral arterial disease) and additional risk factors placing them at increased cardiovascular risk.5 Patients were required to have an LDL-C level of at least 70 mg/dL or non–high-density lipoprotein cholesterol levels of 100 mg/dL while taking an optimized lipid-lowering regimen including a high-intensity or moderate-intensity statin (to convert cholesterol levels to millimoles per liter, multiply by 0.0259)[16,17]. Patients were randomized 1 to 1 to receive subcutaneous evolocumab (either 140 mg every 2 weeks or 420 mg monthly, per patient preference) or matching placebo injection. Patients were followed up for a median of 2.2 years (interquartile range [IQR], 1.8-2.5 years; maximum, 3.6 years). Ethics committee approvals for the FOURIER trial were obtained from all relevant organizations locally or through a central institutional review board within the country. Each patient provided written informed consent.

**DECLARE TIMI** was a randomized, double-blind, multinational, placebo-controlled, phase 3 trial of dapagliflozin in patients with type 2 diabetes and established atherosclerotic cardiovascular disease or multiple risk factors for atherosclerotic cardiovascular disease[18]. Eligible patients were 40 years of age or older and had type 2 diabetes, a glycated hemoglobin level of at least 6.5% but less than 12.0%, and a creatinine clearance of 60 ml or more per minute. Eligible patients also had multiple risk factors for atherosclerotic cardiovascular disease or had established atherosclerotic cardiovascular disease (defined as clinically evident ischemic heart disease, ischemic cerebrovascular disease, or

peripheral artery disease). Participants with multiple risk factors were men 55 years of age or older or women 60 years of age or older who had one or more traditional risk factors, including hypertension, dyslipidemia (defined as a low-density lipoprotein cholesterol level >130 mg per deciliter [3.36 mmol per liter] or the use of lipid-lowering therapies), or use of tobacco

**TMDU study** was a historical cohort study that enrolled 1,963 AF patients who underwent PV isolation and catheter ablation for the first time at TMDU and its associated hospitals. At enrollment, we obtained past medical history and the information of lifestyle. All the specimens were extracted from white blood cells. Patients were followed along with serial correction of AF outcome. Genomic DNAs of control subjects were provided by JCRB Japanese-origin B cell lines and DNA bank, National Institutes of Biomedical Innovation, Health and Nutrition.

**Beat-AF** and **Swiss-AF** are two ongoing prospective observational cohort studies of patients with established AF. AF patients were recruited from 14 sites across all major linguistic regions in Switzerland between 2010 and 2017. Main inclusion criteria were documented AF and age >65 years, although a small convenience sample of patients aged <65 years was also recruited. Patients with an acute illness in the last 4 weeks and those unable to provide informed consent were excluded. Control patients were obtained from the GAPP study, which is a population based cohort study of initially healthy individuals in the Principality of Liechtenstein. Patients with any preexisting major illness were excluded (including patients with known AF). Each participant of all studies provided written informed consent.

Study participants in the **UWO-UBC AF** cohort were recruited from the referral bases for AF management at the London Health Sciences Centre, London, Ontario, Canada and St. Paul's Hospital, Vancouver, British Columbia, Canada. All study participants had at least one episode of electrocardiographically documented AF characterized by erratic atrial activity without distinct P waves and irregularly irregular QRS intervals lasting >30 seconds. Enrollment required that study participants had undergone, at minimum, a clinical history, physical examination, 12-lead ECG, and echocardiogram. The study participants from London were AF cases that had been scheduled to undergo an atrial fibrillation ablation procedure. Study participants from Vancouver had early onset "lone" AF defined as development of the arrhythmia in the absence of known clinical risk factors prior to 60 years of age. These risk factors included hypertension, coronary artery disease, left ventricular ejection fraction <50%, moderate to severe valvular heart disease, hyperthyroidism, obstructive sleep

apnea, and presence of a known underlying inherited channelopathy or cardiomyopathy. All study participants provided informed written consent under protocols that were approved by the research ethics boards of Western University and the University of British Columbia.

# Banner author contributions

### Regeneron Genetics Center

All authors/contributors are listed in alphabetical order.

RGC Management and Leadership Team
Goncalo Abecasis, Aris Baras, Michael Cantor, Giovanni Coppola, Andrew Deubler, Aris Economides, Luca A. Lotta, John D. Overton, Jeffrey G. Reid, Alan Shuldiner, Katia Karalis and Katherine Siminovitch
Contribution: All authors contributed to securing funding, study design and oversight. All authors reviewed the final version of the manuscript.

Sequencing and Lab Operations
Christina Beechert, Caitlin Forsythe, Erin D. Fuller, Zhenhua Gu, Michael Lattari, Alexander Lopez, John D. Overton, Thomas D. Schleicher, Maria Sotiropoulos Padilla, Louis Widom, Sarah E. Wolf, Manasi Pradhan, Kia Manoochehri, Ricardo H. Ulloa.
Contribution: C.B., C.F., A.L., and J.D.O. performed and are responsible for sample genotyping. C.B, C.F., E.D.F., M.L., M.S.P., L.W., S.E.W., A.L., and J.D.O. performed and are responsible for exome sequencing. T.D.S., Z.G., A.L., and J.D.O. conceived and are responsible for laboratory automation. M.S.P., K.M., R.U., and J.D.O are responsible for sample tracking and the library information management system.

Genome Informatics
Xiaodong Bai, Suganthi Balasubramanian, Boris Boutkov, Gisu Eom, Lukas Habegger, Alicia Hawes, Shareef Khalid, Olga Krasheninina, Rouel Lanche, Adam J. Mansfield, Evan K. Maxwell, Mona Nafde, Sean O'Keeffe, Max Orelus, Razvan Panea, Tommy Polanco, Ayesha Rasool, Jeffrey G. Reid, William Salerno, Jeffrey C. Staples,
Contribution: X.B., A.H., O.K., A.M., S.O., R.P., T.P., A.R., W.S. and J.G.R. performed and are responsible for the compute logistics, analysis and infrastructure needed to produce exome and genotype data. G.E., M.O., M.N. and J.G.R. provided compute infrastructure development and operational support. S.B., S.K., and J.G.R. provide variant and gene annotations and their functional interpretation of variants. E.M., J.S., R.L., B.B., A.B., L.H., J.G.R. conceived and are responsible for creating, developing, and deploying analysis platforms and computational methods for analyzing genomic data.

Clinical Informatics:
Michael Cantor and Dadong Li
Contribution: All authors contributed to the clinical informatics of the project

Translational Genetics:
Niek Verweij, Jonas Nielsen, Tanima De and Manuel A. R. Ferreira
Contribution: All authors contributed to the review process for the final version of the manuscript.

Research Program Management
Marcus B. Jones, Jason Mighty and Lyndon J. Mitnaul
Contribution: All authors contributed to the management and coordination of all research activities, planning and execution. All authors contributed to the review process for the final version of the manuscript.

# Supplementary Acknowledgements

**Acknowledgement of TOPMed Studies**

Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). Genome sequencing for “NHLBI TOPMed - NHGRI CCDG: Atherosclerosis Risk in Communities (ARIC)” (phs001211.v4.p3) was performed at the Broad Institute Genomics Platform and Baylor College of Medicine Human Genome Sequencing Center (3U54HG003273-12S2 / HHSN268201500015C, 3R01HL092577-06S1). Genome sequencing for “NHLBI TOPMed - NHGRI CCDG: The BioMe Biobank at Mount Sinai” (phs001644.v1.p1) was performed at the Baylor College of Medicine Human Genome Sequencing Center and McDonnell Genome Institute (HHSN268201600033I, HHSN268201600037I). Genome sequencing for “NHLBI TOPMed - NHGRI CCDG: The Vanderbilt University BioVU Atrial Fibrillation Genetics Study” (phs001624.v2.p2) was performed at the Baylor and McDonnell Genome Institute Platform (3UM1HG008853-01S2, 3UM1HG008898-01S3). Genome sequencing for “NHLBI TOPMed: Cleveland Clinic Atrial Fibrillation (CCAF) Study” (phs001189.v4.p1) was performed at the Broad Institute Genomics Platform (3R01HL092577-06S1). Genome sequencing for “NHLBI TOPMed: The Cleveland Family Study (CFS)” ( phs000954.v4.p2) was performed at the Northwest Genomics Center (HHSN268201600032I, 3R01HL098433-05S1). Genome sequencing for “NHLBI TOPMed: Trans-Omics for Precision Medicine (TOPMed) Whole Genome Sequencing Project: Cardiovascular Health Study” (phs001368.v3.p2) was performed at the Broad Institute Genomics Platform and Baylor College of Medicine Human Genome Sequencing Center (HHSN268201600034I, HHSN268201600033I). Genome sequencing for “NHLBI TOPMed: Genetic Epidemiology of COPD (COPDGene)” (phs000951.v4.p4) was performed at the Broad Institute Genomics Platform and Northwest Genomics Center (HHSN268201500014C, 3R01HL089856-08S1). Genome sequencing for “NHLBI TOPMed: Australian Familial Atrial Fibrillation Study “ (phs001435.v2.p1) was performed at the Broad Institute Genomics Platform (3U54HG003067-12S2 / 3U54HG003067-13S1). Genome sequencing for “NHLBI TOPMed - NHGRI CCDG: Early-onset Atrial Fibrillation in the CATHeterization GENetics (CATHGEN) Cohort” (phs001600.v2.p2) was performed at the McDonnell Genome Institute and Broad Institute Genomics Platform (3UM1HG008853-01S2, 3UM1HG008895-01S2). Genome sequencing for “NHLBI TOPMed: Genomic Activities such as Whole Genome Sequencing and Related Phenotypes in the Framingham Heart Study” (phs000974.v4.p3) was performed at the Broad Institute Genomics Platform (HHSN268201600034I, 3R01HL092577-06S1). Genome sequencing for “NHLBI TOPMed: Heart and Vascular Health Study (HVH)” (phs000993.v5.p2) was performed at the

Broad Institute Genomics Platform and Baylor College of Medicine Human Genome Sequencing Center (3R01HL092577-06S1, 3U54HG003273-12S2 / HHSN268201500015C). Genome sequencing for “NHLBI TOPMed - NHGRI CCDG: The Johns Hopkins University School of Medicine Atrial Fibrillation Genetics Study” (phs001598.v2.p1) was performed at the Broad Institute Genomics Platform (3UM1HG008895-01S2). Genome sequencing for “NHLBI TOPMed - NHGRI CCDG: Massachusetts General Hospital (MGH) Atrial Fibrillation Study” (phs001062.v5.p2) was performed at the Broad Institute Genomics Platform (3U54HG003067-12S2 / 3U54HG003067-13S1; 3U54HG003067-12S2 / 3U54HG003067-13S1; 3UM1HG008895-01S2, 3R01HL092577-06S1). Genome sequencing for “NHLBI TOPMed - NHGRI CCDG: Malmo Preventive Project (MPP)” (phs001544.v2.p1) was performed at the Broad Institute Genomics Platform (3UM1HG008895-01S2). Genome sequencing for “NHLBI TOPMed - NHGRI CCDG: AF Biobank LMU in the context of the MED Biobank LMU” (phs001543.v2.p1) was performed at the Broad Institute Genomics Platform (3UM1HG008895-01S2; HHSN268201500014C). Genome sequencing for “NHLBI TOPMed: Partners HealthCare Biobank (phs001024.v5.p1) was performed at the Broad Institute Genomics Platform” (3R01HL092577-06S1). Genome sequencing for “NHLBI TOPMed - NHGRI CCDG: Penn Medicine BioBank Early Onset Atrial Fibrillation Study” (phs001601.v2.p2) was performed at the Broad Institute Genomics Platform and McDonnell Genome Institute (3UM1HG008853-01S2, 3UM1HG008895-01S2). Genome sequencing for “NHLBI TOPMed: Determining the Association of Chromosomal Variants with Non-PV Triggers and Ablation-Outcome in AF (DECAF)” (phs001546.v2.p1) was performed at the Broad Institute Genomics Platform (3U54HG003067-12S2 / 3U54HG003067-13S1). Genome sequencing for “NHLBI TOPMed - NHGRI CCDG: UCSF Atrial Fibrillation Study” (phs001933.v1.p1) was performed at the Broad Institute Genomics Platform (3UM1HG008895-01S2). Genome sequencing for “NHLBI TOPMed: Defining the Time-Dependent Genetic and Transcriptomic Responses to Cardiac Injury Among Patients with Arrhythmias” (phs001434.v2.p1) was performed at the Broad Institute Genomics Platform (3U54HG003067-12S2 / 3U54HG003067-13S1). Genome sequencing for “NHLBI TOPMed: The Vanderbilt Atrial Fibrillation Registry (VU_AF)” (phs001032.v6.p2) was performed at the Broad Institute Genomics Platform (3R01HL092577-06S1). Genome sequencing for “NHLBI TOPMed - NHGRI CCDG: The Vanderbilt AF Ablation Registry” (phs000997.v5.p2) was performed at the Broad Institute Genomics Platform (3U54HG003067-12S2 / 3U54HG003067-13S1; 3UM1HG008895-01S2; 3UM1HG008895-01S2, 3R01HL092577-06S1). Genome sequencing for “NHLBI TOPMed: Novel Risk Factors for the Development of Atrial Fibrillation in Women” (phs001040.v5.p1) was performed at the Broad Institute Genomics Platform (3R01HL092577-06S1). Genome sequencing for “NHLBI TOPMed - NHGRI

CCDG: The GENetics in Atrial Fibrillation (GENAF) Study" (phs001547.v2.p1) was performed at the Broad Institute Genomics Platform (3UM1HG008895-01S2). Genome sequencing for "NHLBI TOPMed - NHGRI CCDG: Intermountain INSPIRE Registry" (phs001545.v2.p1) was performed at the Broad Institute Genomics Platform (3UM1HG008895-01S2). Genome sequencing for "NHLBI TOPMed: The Jackson Heart Study" (phs000964) was performed at the Northwest Genomics Center (HHSN268201100037C). Core support including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

The Atherosclerosis Risk in Communities study has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health, and Human Services, under Contract nos. (75N92022D00001, 75N92022D00002, 75N92022D00003, 75N92022D00004, 75N92022D00005). The authors thank the staff and participants of the ARIC study for their important contributions.

The CHS research was supported by NHLBI contracts HHSN268201200036C, HHSN268200800007C, HHSN268201800001C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, 75N92021D00006; and NHLBI grants U01HL080295, R01HL087652, R01HL105756, R01HL103612, R01HL120393, R01HL172803 and U01HL130114 with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided through R01AG023629 from the National Institute on Aging (NIA). A full list of principal CHS investigators and institutions can be found at CHS-NHLBI.org. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health

The HVH Study was supported by grants HL068986, HL085251, HL095080, and HL073410 from the National Heart, Lung, and Blood Institute.

The GENAF study is funded by the Norwegian Research Council with a Mobility Grant (240149) and Young Research Talent grant (287086); the South-Eastern Health Authorities with a PhD-grant (2019122); Vestre Viken Hospital Trust with a PhD-grant; afib.no - the Norwegian Atrial Fibrillation Research Network; "Indremedisinsk Forskningsfond" at Bærum Hospital.

The GGAF study is supported by funding to the 5 sources that form GGAF. The AF RISK study is supported by the Netherlands Heart Foundation (grant NHS2010B233), and the Center for Translational Molecular Medicine. Both the Young-AF and Biomarker-AF studies are supported by funding from the University Medical Center Groningen. The GIPS-III trial was supported by grant 95103007 from ZonMw, the Netherlands Organization for Health Research and Development. The PREVEND study is supported by the Dutch Kidney Foundation (grant E0.13) and the Netherlands Heart Foundation (grant NHS2010B280).

The Jackson Heart Study is supported and conducted in collaboration with Tougaloo College (75N92025D00038), Jackson State University (75N92025D00039), University of Southern Mississippi (75N92025D00040), G.A. Carmichael Family Health Center (75N92025D00041), Wake Forest University Health Sciences (75N92025D00036), and the University of Mississippi Medical Center (75N92025D00037) contracts from the National Heart Lung and Blood Institute (NHLBI) with additional support from the National Institute of Minority Health and Health Disparities (NIMHD). The authors also wish to thank the staffs and participants of the JHS. The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the U.S. Department of Health and Human Services.

The WGHS is supported by the National Heart, Lung, and Blood Institute (HL043851 and HL080467) and the National Cancer Institute (CA047988 and UM1CA182913) with funding for genotyping provided by Amgen; AF endpoint confirmation was supported by HL-093613 and HL116690 and a grant from the Harris Family and Watkin’s Foundation.

COPDGene The project described was supported by Award Number U01 HL089897 and Award Number U01 HL089856 from the National Heart, Lung, and Blood Institute. COPDGene is also supported by the COPD Foundation through contributions made to an Industry Advisory Board that has included AstraZeneca, Bayer Pharmaceuticals, Boehringer-Ingelheim, Genentech,

GlaxoSmithKline, Novartis, Pfizer, and Sunovion. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health.

The TIMI Study Group at Brigham and Women's Hospital is supported by the institutional research grant from Anthos, AstraZeneca, Boehringer Ingelheim, Daiichi Sankyo, Janssen, National Institutes of Health, Novartis. Consultancies with Anthos, Bayer, Bristol Myers Squibb, Boehringer Ingelheim, Daiichi Sankyo, Janssen, Pfizer. The TIMI Study Group has received institutional research grant support through Brigham and Women's Hospital from Abbott, Amgen, Anthos Therapeutics, ARCA Biopharma, Inc., AstraZeneca, Daiichi-Sankyo, Eisai, Intarcia, Ionis Pharmaceuticals, Inc., MedImmune, Merck, Novartis, Pfizer, Regeneron Pharmaceuticals, Inc., Roche, The Medicines Company, Zora Biosciences.

**Genomics England Acknowledgements**

We gratefully acknowledge the participants of the National Genomic Research Library (NGRL), whose contributions made this research possible. Secure access to the NGRL under project ID RR1026 "Elucidating the rare genetic architecture of early-onset atrial fibrillation" was provided by Genomics England, which delivers the NGRL in partnership with NHS England, and is wholly owned by the UK Department of Health and Social Care. The NGRL contains participants' health data collected by the NHS as part of their care, along with samples and data from their participation in research, for which fully informed consent has been obtained. This includes genomic and clinical data provided through the NHS Genomic Medicine Service, as well as data obtained through research studies, including the 100,000 Genomes Project and the Generation Study, both of which are delivered in partnership with the NHS, and from other research cohorts involving external collaborators.

# Supplementary Methods

### Sequencing protocols and quality controls

*Whole-genome sequencing data from TOPMed-CCDG*

We used the same TOPMed dataset as described in our previous publication. We utilized the TOPMed whole-genome sequencing freeze 8 database. The sequencing methods are described in https://topmed.nhlbi.nih.gov/topmed-whole-genome-sequencing-methods-freeze-8 and in Taliun et al.[19] In brief, whole-genome sequencing was performed at 7 different sequencing centers, after which the Information Research Center centrally harmonized read aliments, discovered variants, and called genotypes. This freeze included 137,977 whole-genome sequenced samples, of which we first selected 126,572 samples that had IRC QC metrics available. Then, we removed samples that failed IRC quality controls or did not have AF status available, leaving 37,383 individuals. We additionally removed genetically-determined duplicates (N = 44), sex-mismatched samples (N = 30), samples with call rate ≤ 95% (N = 0), outliers (N = 5) and mislabeled samples (N = 1). To determine sex-mismatched samples, we selected high quality (MAF> 5% and call rate > 0.99) variants on the X chromosomal non-PAR region. Then we pruned variants using the PLINK2 *–indep-pairwise 50 5 0.2* option[20]. We removed 1) males if their F statistics was <0.75 and 2) females if their F statistics was >0.55. Samples were determined to be outlier samples if they were outside of 8 standard deviations from the mean for any of several additional metrics, including the Ti/Tv ratio, het/hom ratio, SNV/INDEL ratio, and the number of singletons. In the final sample set, 15,770, 13,857, 5,049, and 2,627 samples were sequenced in Baylor Human Genome Sequencing Center, Broad Institute of MIT and Harvard, University of Washington Northwest Genomics Center, McDonnell Genome Institute, respectively. We used ADMIXTURE (ref.[21]) to assign the ancestries to TOPMed-CCDG samples using the 1000 Genomes (1KG) dataset as a reference panel[22]. To this end, we used high-quality variants present in both TOPMed-CCDG and 1KG, with MAF >5% in 1KG, missingness <1% in both TOPMed-CCDG and 1KG, and consecutively pruned in each 1KG super-population with *–indep-pairwise 50 5 0.2* in PLINK2 (ref.[20]). Ancestry probabilities were learned from the 1KG data and projected onto TOPMed-CCDG samples with ADMIXTURE. When the estimated probability for a sample was >80% for a given super-population, they were assigned to that ancestry. In this process, 6,425 samples were assigned to African American, 363 samples to Admixed American, 209 samples to East Asian, 24,702 samples to European, and 207 samples to South Asian ancestry. We found

5,397 samples with probability <80% for all continental populations, who were assigned to undetermined ancestry.

We then performed variant-level quality-control procedures within the samples with available AF status (N = 329,976,117 variants). First, we removed variants in low-complexity regions (N = 383,344) and variants with call rate <95% (N = 512,688). We then performed Hardy-Weinberg equilibrium tests within several homogeneous ancestral groups (European, South Asian, East Asian, Admixed American and African). We utilized a *P*-value threshold of $1x10^{-6}$ and filtered a variant if 1) the null hypothesis was rejected in at least two homogeneous populations or 2) the null hypothesis was rejected in one homogeneous population with a MAF>1% (N = 8,385). Finally, after removing samples from sample QC described above, the remaining monomorphic variants were removed (N = 166,542). This procedure left 328,905,158 variants for analysis in TOPMed-CCDG.

The Data Coordinating Center form TOPMed generated the principal components and a sparse-kinship matrix using KING (ref.[23]), PCAiR (ref.[24]) and PCRelate (ref.[25]), which has been made available to TOPMed researchers (https://topmed.nhlbi.nih.gov/group/dcc).

*Whole-exome sequencing data from Geisinger MyCode*

Whole-exome sequencing was performed in collaboration with Regeneron Genetics Center as previously described[6,26,27]. In brief, a modified version of the xGEN probe from Integrated DNA Technologies (IDT) was used for target sequence capture. Sequencing was performed using 75 base pair paired-end reads using Illumina HiSeq 2500 or Illumina NovaSeq instruments. 90% coverage of the targeted bases at 20x in 99% of the samples was achieved. Reads were aligned to the GRCh38 human reference genome using BWA-mem. Variant calling was performed using WeCall. A cohort-level VCF was then created using GLnexus for downstream analysis. Sample-level calls with depth of coverage at a particular site (DP<7 for SNPs, DP<10 for indels) were changed to no-calls. Variant sites were required to have at least one sample meeting an alternate allele balance threshold (AB>=0.15 for SNPs, AB>=0.20 for indels).

*Whole-exome sequencing data from CCDG-WES, FOURIER and DECLARE*

We analyzed a joint-call set of 110,061 raw (pre-QC) exome sequences, which included data from CCDG-WES samples (including an initial freeze of MGB) and samples from the FOURIER trial. We also analyzed a separate joint-call set of 13,279 raw exome sequences of samples from the

DECLARE trial. Whole-exome sequencing of CCDG-WES and DECLARE samples was performed at the Broad Institute of MIT and Harvard from 2019 to 2022. Samples were sequenced on the Illumina NovaSeq 6000 sequencing system using TWIST human exome capture kits. The sequencing process included sample prep (385bp shear + Kapa Hyperprep LC + IDT/Broad Indices), in-solution hybrid capture (with a custom exome panel manufactured by TWIST, ~37Mb target), sequencing (Illumina, 150bp paired reads) and standard identification quality-control checks. Hybrid selection libraries typically meet or exceed 85% of targets at 20x, comparable to ~55x mean coverage. Raw reads were aligned to GRCh38 using BWA and duplicate-marked and filtered using Picard, after which per-sample gVCF files were created using GATK, all according to the GATK best practices guidelines[28]. For the FOURIER dataset, samples have previously been whole-exome sequenced at deCODE in Iceland, in 2017. Samples were sequenced on Illumina HiSeq X machines. Alignment to GRCh38 was performed with BWA v0.7.10, after which BAM files were created using fastq2BAM v1.8.9, producing SAM specification v1.4 BAM files. BAM files were then reprocessed to CRAM files at the Broad Institute of MIT and Harvard, by unmapping of BAM files and subsequent remapping using SAMtools according to GATK best practices guidelines[28], after which per-sample gVCF files were produced using GATK. Per-sample gVCF files from all CCDG-WES and FOURIER samples were combined and processed at the Broad Institute of MIT and Harvard for joint-genotyping of small SNVs and INDELs, using GATK according to GATK best practices guidelines[28]. DECLARE samples were separately joint-called using similar methods.

After genotype calling, we split CCDG-WES and FOURIER samples, since these cohorts were based on different sequencing platforms and different timepoints, and therefore we had three distinct call sets in this pipeline (CCDG-WES, FOURIER, DECLARE). We then executed genotype, variant, and sample QCs within these datasets, separately, using the same methods. We used hail v0.2 (https://hail.is/docs/0.2/index.html) to split multi-allelic variants to represent separate biallelic sites, and then applied genotype hard-filters. We removed genotypes if they met the following criteria:

- For homozygous reference calls: Genotype Quality < 20; Genotype Depth < 10; Genotype Depth > 200
- For heterozygous calls: (A1 Depth + A2 Depth)/Total Depth < 0.9; A2 Depth/Total Depth < 0.2; Genotype Quality < 20; Genotype Depth < 10; Genotype Depth > 200
- For homozygous alternative calls: (A1 Depth + A2 Depth)/Total Depth < 0.9; A2 Depth/Total Depth < 0.9; Genotype Quality < 20; Genotype Depth < 10; Genotype Depth > 200

In initial variant-level quality control procedures, variants were removed if they were assigned a “FAIL” status by the GATK Variant Quality Score Recalibration (ref.[29]) machine learning algorithm, had call rates < 90% (not applied to sex-chromosomal variants), were monomorphic, failed a stringent Hardy-Weinberg Equilibrium test at *P*-value <$1x10^{-15}$ (applied only in DECLARE; not applied to sex-chromosomal variants), or were located within Ensembl low-complexity regions.

We then used ADMIXTURE (ref.[21]) to assign ancestral probabilities to samples for 5 major super-populations (European, East Asian, South Asian, African and Admixed American), using the 1KG data as a training set[22]. To this end, we used autosomal variants from the 1KG dataset that had MAF>1% in at least one ancestral super-population and passed the Hardy-Weinberg Equilibrium test in each ancestry (*P*<$1x10^{-6}$). Variants were then consecutively pruned in each ancestral group using *–indep-pairwise 50 5 0.2* in PLINK2 (ref.[20]). ADMIXTURE was then used to learn the probabilities from the 1KG data, and to subsequently project ancestry probabilities onto samples in the exome sequencing datasets. Individuals with probability >80% for an ancestry were assigned to that super-population.

We then performed principal component analysis (PCA) and kinship inference. Utilizing the same variant subset described above for the ADMIXTURE analysis, we first ran KING to compute pair-wise kinship estimates[23]. We then ran a randomized algorithm[30] implemented in PCAir to compute principal components of genetic ancestry[24,30]. In the process, PCAir was fed the KING kinship estimates to determine the largest unrelated subset of the dataset in which to perform PCA (using a cutoff of $2^{-9/2}$), after which PCs were projected on the remaining related samples.

Then we applied sample-level filtering. We first removed individuals with overall call rates < 90%. We then removed individuals who were outside 8 standard deviations from the mean for several PC-adjusted metrics, namely Ti/Tv ratio, het/hom ratio, SNV/INDEL ratio, and the number of singletons (all normalized by regressing the first 10 PCs of ancestry out of the metrics using linear regression). We computed pair-wise heterozygote concordance rates with KING (ref.[23]) version 2.2.5 (using the same set of autosomal variants described above for the ADMIXTURE analysis); sample duplicates were identified using a cutoff of >80% and were removed from the dataset. We removed samples without available AF phenotype status; for AF cases from TMDU, we did not have age and sex data available, and therefore we imputed age for those samples to the case mean (68.4). We then used PLINK2 to identify sex-mismatched samples, by computing X-chromosomal inbreeding

coefficients (F) for each sample (using non-PAR X-chromosomal variants with MAF>0.5%, missingness <5%, Hardy-Weinberg equilibrium test $P>1x10^{-6}$ in each ancestry and consecutively pruned in each ancestral group with *–indep-pairwise 50 5 0.2* in PLINK2); genetically-determined female sex was assigned at F<0.5, while genetically-determined male sex was assigned at F>0.75; samples with blatant mismatches between self-reported and genetically-determined sex were removed from the datasets. Finally, for datasets where we could analyze individual-level data in one compute environment (CCDG-WES, FOURIER, DECLARE, MGB, TOPMed-WGS), we then removed individuals who were duplicated or closely related to individuals from other datasets (methods described below).

*Whole-genome sequencing data from the UK Biobank*

We used the WGS data released to the DNAnexus platform. To this end, we used the pVCF files released in 2024, processed through the DRAGEN pipeline (https://www.ukbiobank.ac.uk/media/c3zpw015/uk-biobank-final-whole-genome-sequencing-release-faqs_v3.pdf). We processed the >150,000 pVCF files through various steps, which included genotype-level, variant-level, and sample-level QC procedures. After central filtering, we identified WGS data for 490,542 samples, which entered the pipeline described below.

First, we processed the pVCF files by performing genotype-level refinement, and subsequent trimming of the VCF meta-data to reduce file size. To this end, *bcftools* (v1.16) with the *setGT* plugin (v1.20) was used to set low-quality genotypes to missing; genotypes flagged as 'lowDepth' or that failed the genotype-level DRAGEN machine-learning algorithm were set to missing. For details, see https://github.com/seanjosephjurgens/vcf_trimmer_genotypeqcer.

Next, we applied variant-level QC on the processed pVCF files. Using *bcftools* (v1.16), we removed variants that failed the variant-level DRAGEN machine-learning algorithm, or failed any other filter (ie INFO=/=PASS), after which multi-allelic variants were split to represent several biallelic variants. For details, see https://github.com/seanjosephjurgens/vcf_splitter_variantqcer. The pVCF files were then merged into larger pVCF files (50-300 per chromosome; see https://github.com/seanjosephjurgens/vcf_merger), after which we converted these pVCFs to PLINK2 format (see https://github.com/seanjosephjurgens/vcf2gds/tree/PLINK2), after which these chunks were merged into a single PLINK2 file per chromosome (see https://github.com/seanjosephjurgens/plink_merger). We then used the *minrep()* function from *hail* (v0.2.116) to normalize variant IDs and to convert the IDs to minimal representation. Finally, we used

PLINK2 (vLinux_avx2_20240818) to remove variants that failed specific filters, including those with >0.1 missingness, those with minor allele count <1, and those with ExcessHet *P*<1e-20 (or Hardy-Weinberg-Equilibrium test *P*<1e-20 for X-chromosome variants). This process resulted in 1,195,517,695 autosomal, 47,666,140 non-autosomal X-chromosal, and 1,626,190 pseudo-autosomal X-chromosomal variants after variant-level QC.

We then defined a filtered genomic dataset, of pruned high-quality autosomal variants, for several downstream sample-QC procedures (using PLINK2 v-linux_x86_64_20240818). This filtered dataset was made by filtering to variants with minor allele frequency >1%, missingness <1%, LD-pruned (--indep-pairwise 500 200 0.1 and --indep-pairwise 2000 400 0.1), excluding long-range LD regions, and finally taking 100k random markers.

Using this filtered dataset, we compute sample relatedness, ancestry assignments, and to perform principal component analysis for the UKB samples: First, *king2* (v2.3.2) was used to estimate heterozygote concordance rates and kinship coefficients for each sample-sample combination. Then, *flashPCA* (v2.0) was used to perform principal component analysis (PCA) on unrelated (kinship<0.042) samples, with subsequent projection of PCs onto the remaining samples. Ancestry assignments were then performed using *ADMIXTURE* (v.1.3.0), by i) merging the data with the 1KG dataset, ii) sequently pruning variants in each continental superpopulation within the 1KG samples, iii) learning ancestry probabilities from the 1KG samples, and iv) projecting ancestry assignments onto the UKB samples; samples were assigned to a continental ancestry if the probability was >=0.8.

We then applied sample-level QC filtering. We first removed all samples with revoked consents (N=245 removed). We then flagged individuals with ambiguous or mismatched genetically-inferred sex, as compared to their self-reported sex (N=210). To this end, we used PLINK2 (v-linux_x86_64_20240818) to identify sex-mismatched samples, by computing X-chromosomal inbreeding coefficients (F) for each sample (using non-PAR X-chromosomal variants with MAF>0.5%, missingness <1%, pruned using –indep-pairwise 500 200 0.1, and then pruned again using --indep-pairwise 2000 400 0.1); genetically-determined female sex was assigned at F<0.5, while genetically-determined male sex was assigned at F>0.8. We then removed one sample from duplicated pairs (using a heterozygote concordance rate >0.8; N=235 flagged). We then flagged samples that did not pass a centrally-computed metric for DNA quality (N=239 flagged); samples that had a DNA contamination score of >=2 (N=67 flagged); samples that had high missingness (>1%

across all autosomal variants; N=0 flagged); and samples that were outliers for various other metrics, including the PC-adjusted Ti/Tv ratio, PC-adjusted Het/Hom ratio, PC-adjusted SNV/indel ratio, and the PC-adjusted number of singletons (N=1653 flagged). Overall, of the 490,542 initial samples, 490,296 had appropriate consents, of which 487,917 remained after removing all flagged samples.

*Whole-exome sequencing data from MGB*

We described processing and quality-control of exome sequencing data from 53k MGB samples in a previous publication[31]. In essence, the pipeline was similar to the pipeline described above for the CCDG, FOURIER, and DECLARE datasets. In the current study, we used this call set, after which we excluded considerable numbers of samples already included in the CCDG call set. Methods for identification of overlapping samples is described in detail below.

*Whole-genome sequencing data from AllofUs V7*

In our main burden testing discovery analyses, we used genomic data from the V7 release of AllofUs. The quality control applied to the genome data was described in our previous publication[31]. It is known that MGB contributes substantially to the AllofUs dataset. Due to biobank policies, however, it was not possible to use genomic data to remove samples present in both datasets. For this reason, as described in more detail below, we instead removed all samples from Massachusetts from the AllofUs dataset as part of the discovery analyses. This dataset will be referred to as the AllofUs_V7only_noMA dataset in the rest of this documentation.

*Whole-genome sequencing data from AllofUs V8*

In downstream analyses pertaining to heritability and replication, we used the V8 release of AllofUs. The AllofUs Research Program performed extensive genotype-level, variant-level, sample-level QC on the PLINK datasets, as described in the documentation (https://support.researchallofus.org/hc/en-us/articles/29390274413716-All-of-Us-Genomic-Quality-Report), resulting in PLINK format datasets of 414,830 samples. We used the PLINK1.9 call sets - the 'exome' callset for the main genomic analyses of this paper, and the 'ACAF' callset for additional QC procedures described below. Given that genotype-level and variant-level QC was already comprehensive, we only removed variants with missingness>10% or minor allele count <1 from both callsets. Out of 106,982,995 variants in original ACAF files, 83,652,100 variants passed QC. Out of 40,598,788 variants in original exome files, 39,574,746 variants passed QC.

We then applied additional processing and QC on the sample-level. Using *king2* (v2.3.2) and *flashPCA* (v2.0), and using the ACAF dataset, we then performed relatedness inference, duplicate identification, and PCA following the exact same computational procedures as described above for the UKB WGS data. We then i) removed samples flagged by central QC procedures, ii) limited to individuals with XX or XY ploidy as determined by central QC procedures, and iii) removed potentially duplicated samples - preferentially keeping samples from a pair if they had electronic health record data available and otherwise keeping the sample with higher call rate. This procedure left 410,400 samples that passed all genomics-based sample QC, of which 315,536 samples had electronic health record data available.

The entire V8 dataset will be referred to as the AllofUs_V8 dataset in the rest of this documentation; this dataset was used for burden heritability analyses. The V8 dataset, excluding samples included in the V7 dataset and excluding samples from Massachusetts (MA), will be referred to as the AllofUs_V8only_noMA dataset; this last dataset was used for replication analyses (N=6,513 AF cases and N=96,848).

*Overlapping samples across datasets*

Since it was not possible to account for relatedness across datasets contributing to the discovery meta-analysis, false findings could be introduced. For instance, samples from the MGB biobank were included in the MGB callset, the CCDG-WES callset, the TOPMed-CCDG callset, and potentially in the AllofUs callsets[31]. We therefore performed several steps to avoid sample overlap across datasets as much as possible. The best approach to avoid sample overlap would be to combine the individual-level data from all datasets and identify overlapping samples using the genetic data. Unfortunately, many of the individual-level datasets included in this study live within their own protected environments, preventing such an approach.

Therefore, we initially only performed such an individual-level approach for datasets that could be combined in the same analytical environment. To this end, we could identify samples overlapping between the TOPMed-CCDG callset, the CCDG-WES callset, the MGB callset, the FOURIER callset, and the DECLARE callset. To perform relatedness analysis, we first identified autosomal genetic variants shared across all these callsets, then filtered to variants with MAF>0.001 in the TOPMed-CCDG callset (a highly diverse multi-ancestry dataset), and finally pruned in the TOPMed-CCDG dataset using '--indep-pairwise 2000 200 0.1' in PLINK2. Using these autosomal

pruned variants, we then *king* to compute heterozygote concordance rates across datasets, flagging pairs with >0.8 as potential duplicates. Using the resulting list of putative overlapping pairs, we removed one individual from each pair, while preferentially keeping all samples from TOPMed-CCDG (the only WGS dataset). This approach left 87,300 samples for CCDG-WES, 21,933 samples from MGB, 15,991 samples from FOURIER, and 12,410 samples from DECLARE.

Because we previously described overlap between samples from AllofUs and MGB, we also defined subsets of the AllofUs callsets from which we removed samples from Massachusetts, as described above (the AllofUs_V7only_noMA and AllofUs_V8only_noMA datasets). To this end, we used ZIP code and state of residence data to identify individuals living in Massachusetts. Any individuals found to live in Massachusetts or have missing data for these variables were excluded from these restrictive callsets. While it is unlikely that all Massachusetts samples from AllofUs are truly overlapping with the MGB dataset, this approach likely removed the vast majority of overlapping samples. To remain conservative in our analysis, we therefore chose this rather stringent but safe approach.

**Annotation and rare variant burden testing**

In all datasets other than Geisinger MyCode, we used highly similar pipelines to annotate genetic variants in a standardized way. In all these datasets, the Variant Effect Predictor (VEP)[32] with the LOFTEE plugin (https://github.com/konradjk/loftee), as well as the dbNSFP database (https://www.dbnsfp.org/)[33], were used to annotate genetic variants. In the UK Biobank, TOPMed-CCDG, CCDG-WES, MGB, FOURIER, DECLARE, AllofUs_V7, and AllofUs_V8 datasets, we used VEP v105, while in Geisinger MyCode we used VEP v104. In the MGB, AllofUS_V7 and Geisinger MyCode datasets we used dbNSFP v4.3, while in the UK Biobank, TOPMed-CCDG, CCDG-WES, FOURIER, DECLARE and AllofUs_V8 datasets dbNSFP v4.2 was used.

In all above mentioned datasets, we first used VEP to identify rare coding variants affecting the ENSEMBL canonical gene transcripts. From these variants, high-confidence loss-of-function (LOF) variants, including protein-truncating variants and high-impact splice variants, were identified using LOFTEE; we distinguished and analyzed two broad groups of LOF variants, one irrespective of potential flags, and a more stringent group excluding LOFs with any flags (LOFnoflag). Such LOFTEE flags may include variants affecting NAGNAG sites, variants affecting noncanonical splice sites, variants affecting non-conserved exons, or variants affecting single-exon transcripts.

Missense variants were also identified using VEP, and functional bioinformatic-predicted consequences were then inferred using various prediction tools. For the REVEL prediction tool[34], we used the variant prediction associated with the canonical transcript annotated through the dbNSFP database; we considered missense variants with a REVEL score of >=0.7 as predicted-damaging missense-REVEL variants.

We also curated additional, newer bioinformatic prediction tools. We downloaded the AlphaMissense predictions[35] (https://github.com/google-deepmind/alphamissense?tab=readme-ov-file), using the 'AlphaMissense_hg38.tsv.gz' file released in August 2023. We used gencode v45 to map the ENSEMBL transcript IDs to ENSEMBL gene IDs. We found discrepancies between the transcript used by the developers, and the ENSEMBL canonical transcript; as such, we mapped the variant predictions to variants affecting the canonical transcript from the VEP output, using the ENSEMBL gene IDs. We considered missense variants with 'predicted_pathogenic' classification by AlphaMissense as being predicted-damaging missense-AM variants.

We similarly received predictions for the PrimateAI-3D (PAI3D) tool[36], by requesting access in April 2024 (we received the 'PrimateAI-3D_scores.csv.gz' file). As above, we used gencode to map transcripts to ENSEMBL gene IDs and used these to match predictions to canonical gene transcripts from VEP. We considered missense variants with PAI3D score >=0.803 as predicted-damaging missense-PAI3D variants.

We also downloaded the popEVE predictions[37] from December 2023 (file '2023-12-19_popeve_bulk.tar.gz'). Since the popEVE predictions were linked to RefSeq protein IDs, we first mapped the RefSeq protein IDs to ENSEMBL gene IDs (using https://ftp.ncbi.nlm.nih.gov/pub/datasets/command-line/v2/linux-amd64/datasets). In contrast to AlphaMissense and PAI3D, popEVE predictions were on the amino-acid change level, instead of the variant level. We therefore mapped popEVE predictions to missense genetic variants, using the VEP-predicted protein changes. We considered missense variants with popEVE score < -5.056 as predicted-damaging missense-popEVE variants.

**Burden testing models**

In all datasets contributing to our main discovery analysis (UK Biobank, TOPMed-CCDG, CCDG-WES, AllofUs_V7only, Geisinger Mycode, MGB, FOURIER, DECLARE), we performed rare

variant burden testing using various versions of REGENIE. In all datasets, analyses were performed inclusive of all individuals irrespective of ancestry, although sensitivity analyses were also performed restricting to individuals of genetically-inferred European ancestry. In all datasets, step1 regression models were adjusted - at least - for age, sex, and ancestral PCs; step2 models were at least adjusted age, sex, ancestral PCs and the REGENIE polygenic predictor.

**Replication analyses in Genomics England**

The 100,000 Genomes Project is a national UK program, delivered by Genomics England (GeL), that recruited probands with rare diseases and cancer from clinical centers, together with family members, and performed germline and somatic (for a subset of participants with cancer) WGS. We used the Aggregated Variant Calls (AggV2) fileset, containing 76,159 WGS samples and 464,800,405 variants (445,496,423 autosomal). This dataset was sample and variant quality controlled by GeL (as detailed in the GeL Research Environment documentation). We performed a few additional quality control steps after this, as follows. Samples without current consent were removed (n=55) leaving 76,103 samples. Variants with genotype missingness exceeding 5% were excluded, leaving 423,167,597 variants. Variants in low complexity regions were excluded, leaving 408,929,404 variants. Atrial fibrillation cases were defined as individuals possessing ≥ 1 of the billing codes mentioned below. This yielded 3,387 cases and 72,716 controls, where controls were defined as any individuals that were in the 76,103 samples and not a case. The dataset was annotated using VEP 109. A rare variant association analysis was performed on this dataset using the same masking strategy and REGENIE methodology as the discovery cohort.

Billing codes used for case definition in GeL:

| | |
|---|---|
| Atrial fibrillation | SNOMED 49436004 |
| Chronic atrial fibrillation | SNOMED 426749004 |
| Controlled atrial fibrillation | SNOMED 300996004 |
| Paroxysmal atrial fibrillation | HPO HP:0004757 |
| Paroxysmal atrial fibrillation | SNOMED 282825002 |
| Permanent atrial fibrillation | HPO HP:0004754 |
| Rapid atrial fibrillation | SNOMED 314208002 |
| Chronic atrial fibrillation | ICD10 I48.2 |
| Familial atrial fibrillation | SNOMED 715395008 |
| Lone atrial fibrillation | SNOMED 233910005 |

Paroxysmal atrial fibrillation ICD10 I48.0

Permanent atrial fibrillation SNOMED 440028005

Persistent atrial fibrillation SNOMED 440059007

Persistent atrial fibrillation ICD10 I48.1

Preexcited atrial fibrillation SNOMED 762247006

Atrial fibrillation and flutter SNOMED 195080001

Atrial fibrillation and flutter ICD10 I48

Longstanding persistent atrial fibrillation SNOMED 706923002

Non-rheumatic atrial fibrillation SNOMED 233911009

Maze procedure for atrial fibrillation SNOMED 429211003

Hypercoagulable state due to atrial fibrillation SNOMED 129032061000119103

Atrial fibrillation and atrial flutter, unspecified ICD10 I48.9

Atrial fibrillation with rapid ventricular response SNOMED 120041000119109

History of maze procedure for atrial fibrillation SNOMED 429218009

Paroxysmal atrial fibrillation with rapid ventricular response SNOMED 1010405004

Transient cerebral ischaemia due to atrial fibrillation SNOMED 426814001

Atrial flutter HPO HP:0004749

Atrial flutter SNOMED 5370000

Atypical atrial flutter SNOMED 15964901000119107

Atypical atrial flutter ICD10 I48.4

Chronic atrial flutter SNOMED 425615007

Paroxysmal atrial flutter SNOMED 427665004

Typical atrial flutter HPO HP:0031671

Typical atrial flutter SNOMED 720448006

Typical atrial flutter ICD10 I48.3

Atrial fibrillation and flutter SNOMED 195080001

Atrial fibrillation and flutter ICD10 I48

Reverse typical atrial flutter HPO HP:0031672

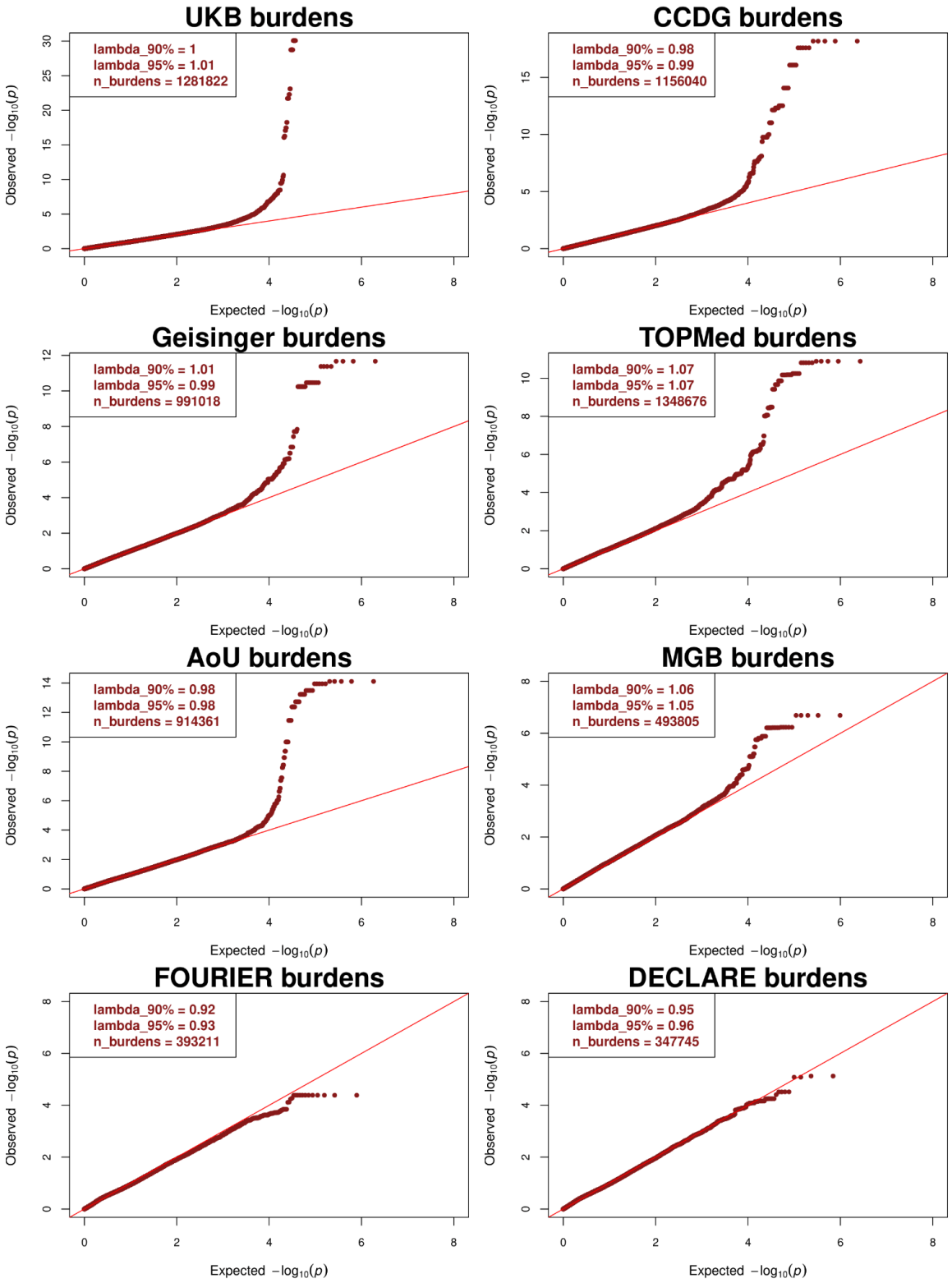
Atrial fibrillation and atrial flutter, unspecified ICD10 I48.9

Percutaneous transluminal ablation of atrial wall for atrial flutter OPCS K622

Percutaneous transluminal ablation of atrial wall for atrial flutter SNOMED 707832008

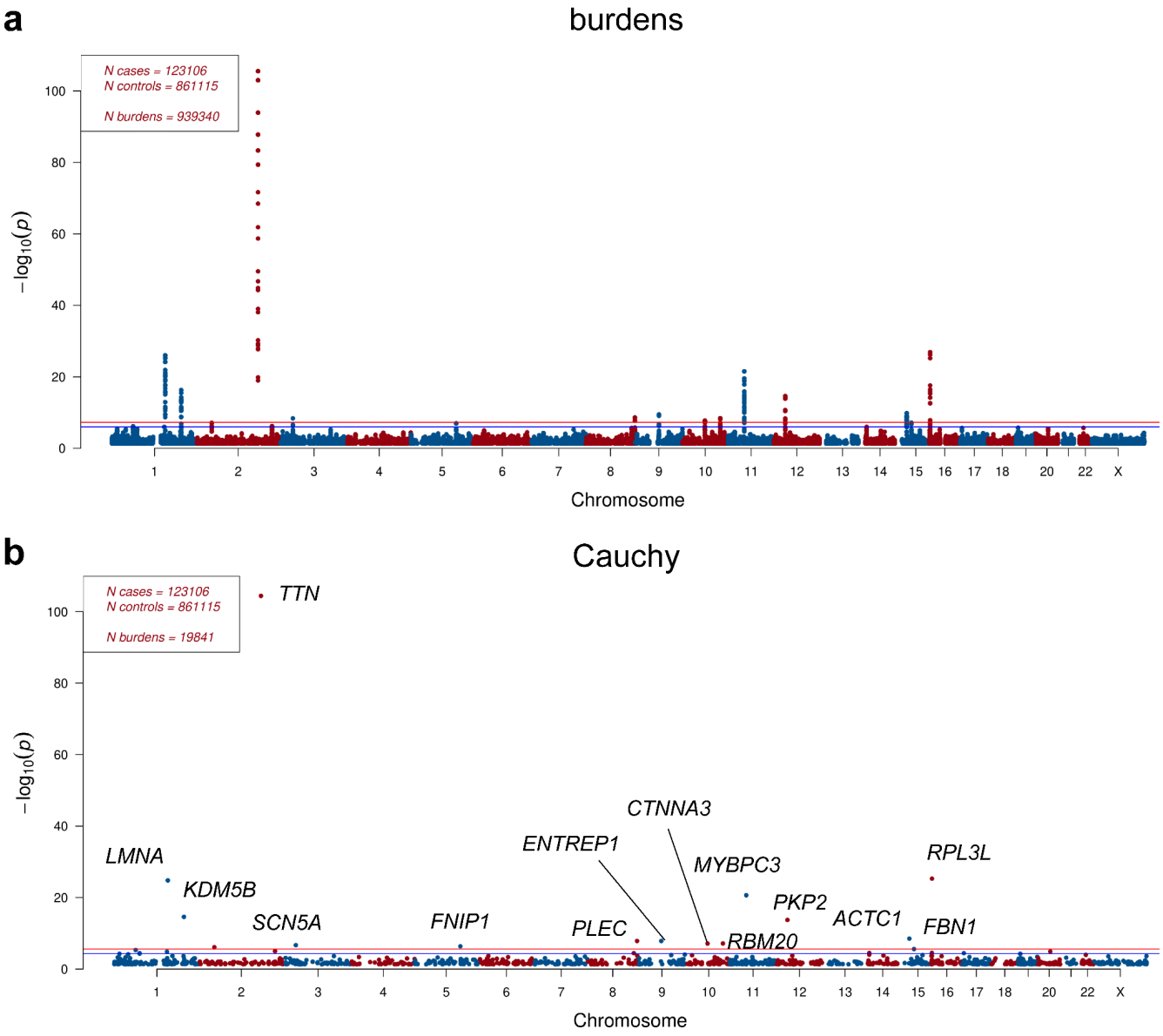
Percutaneous transluminal ablation of conducting system of heart for atrial flutter NEC OPCS K623

# Supplementary Figures

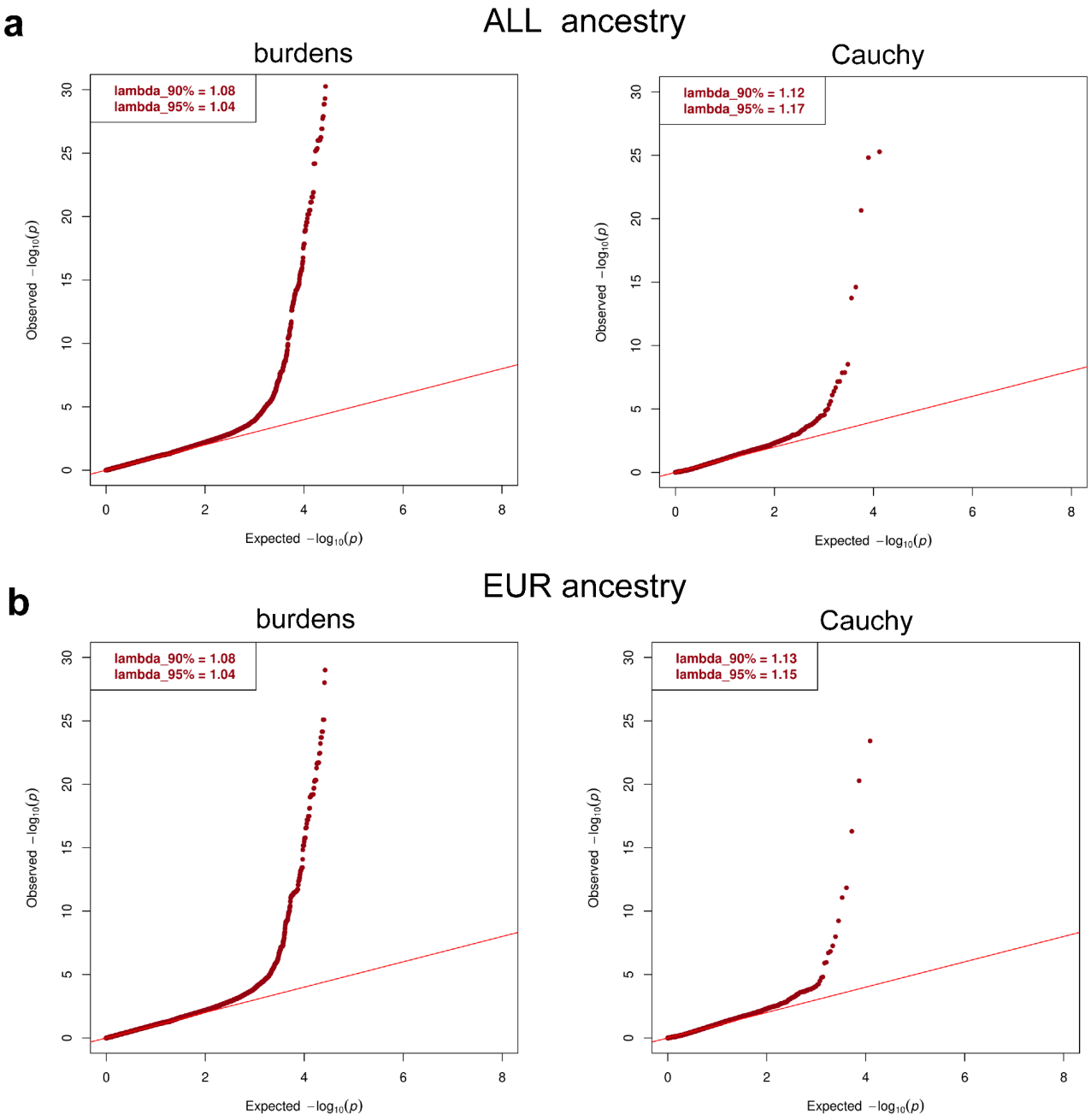



**Supplementary Figure 1: Quantile-quantile plots from mask-based burden testing in each dataset.** Each panel represents a quantile-quantile plot, with expected -log10 *P*-values on the x-axis and observed values on the y-axis; each dot represents burden testing association results for a single

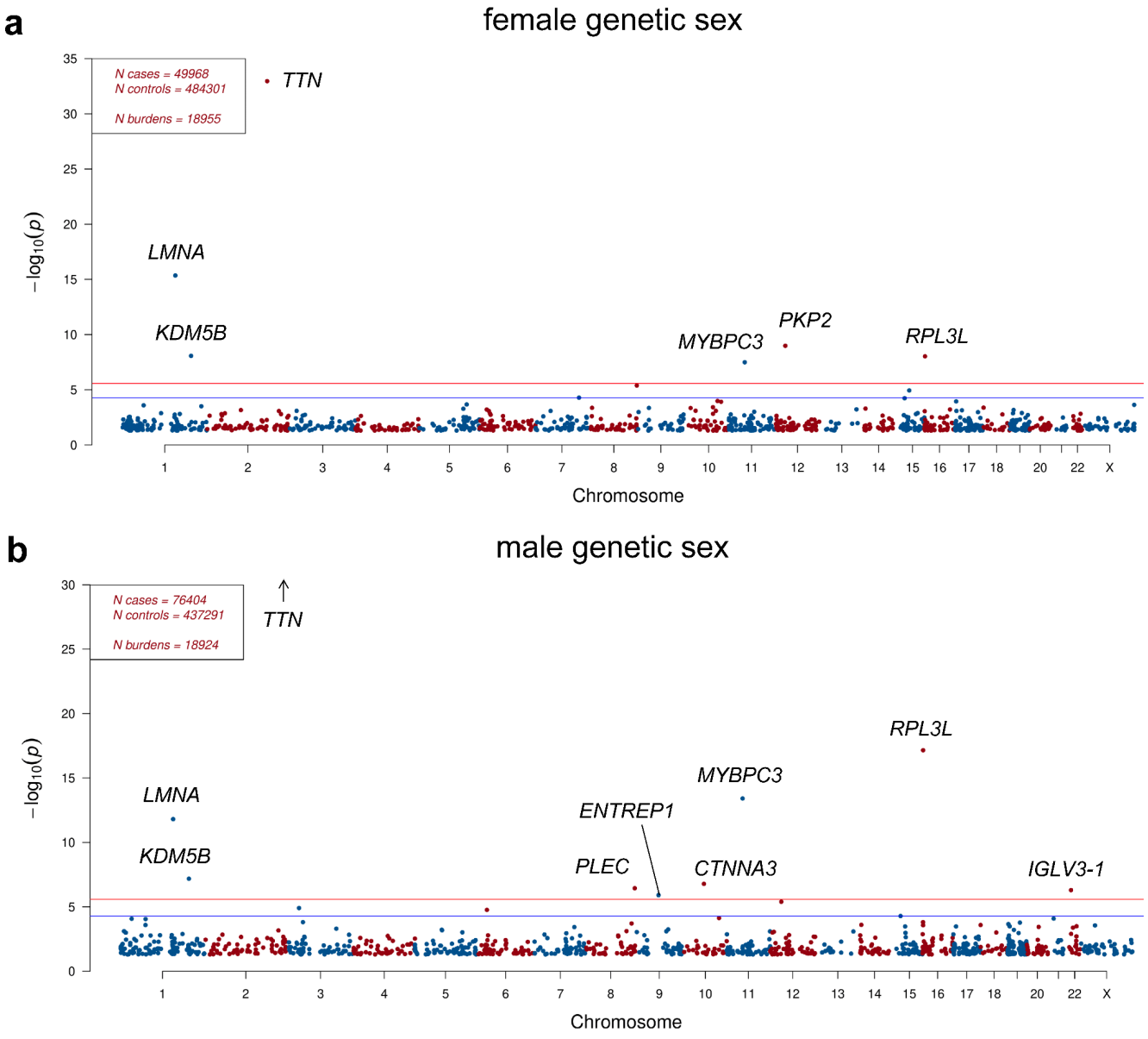
mask (as such, a given gene may be represented up to 60 times). Each panel represents a different analysis dataset. The red lines indicate y=x. Lambda values (representing the observed chi-squared statistic over the expected chi-squared statistic at a given distributional cutoff) are added to the plots, including the lambda value at the 90th and 95th percentile of the test statistic distribution. Overall, good calibration of test statistics was observed across datasets.



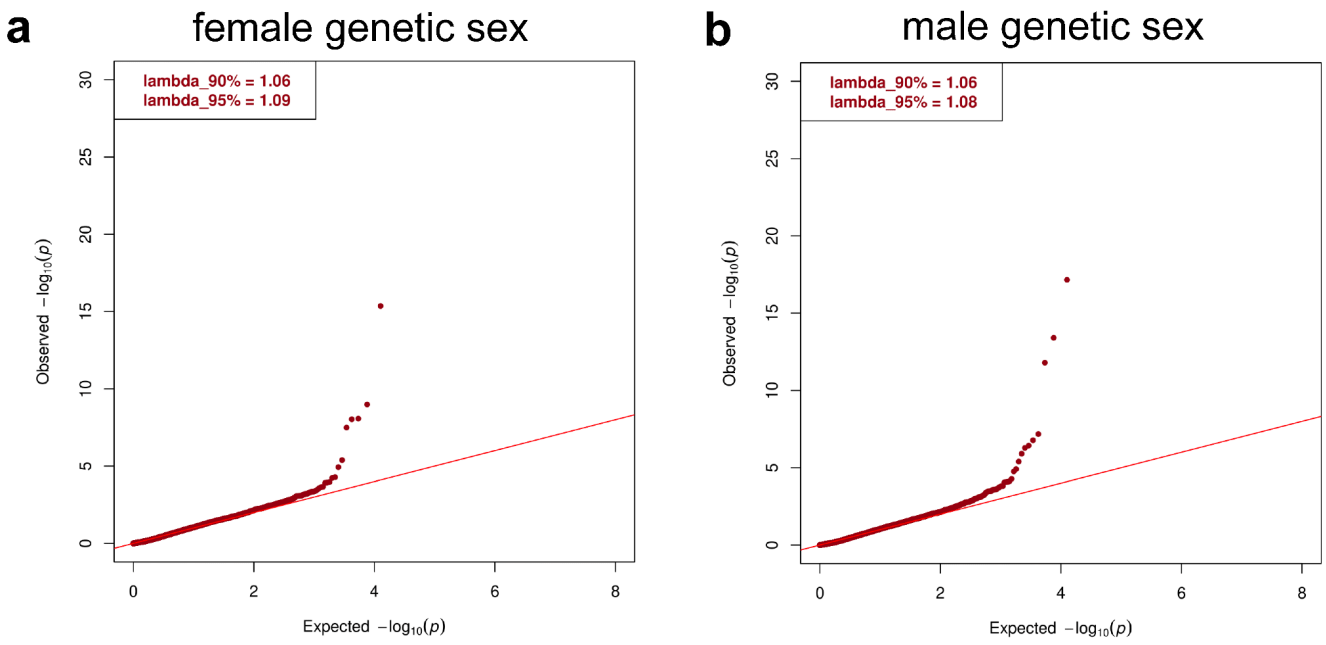

**Supplementary Figure 2: Manhattan plots of the main discovery analysis for mask-based burdens and the gene-based Cauchy combination.** Both panels are Manhattan plots with chromosome position on the x-axis and -log10 *P*-values on the y-axis. Both panels show association results for the main discovery meta-analysis. Panel **a** shows results for mask-based burden testing, with each dot representing a single mask (as such, each gene may be represented up to 60 times). Panel **b** shows results for the Cauchy combination approach, and therefore each dot represents a single gene.



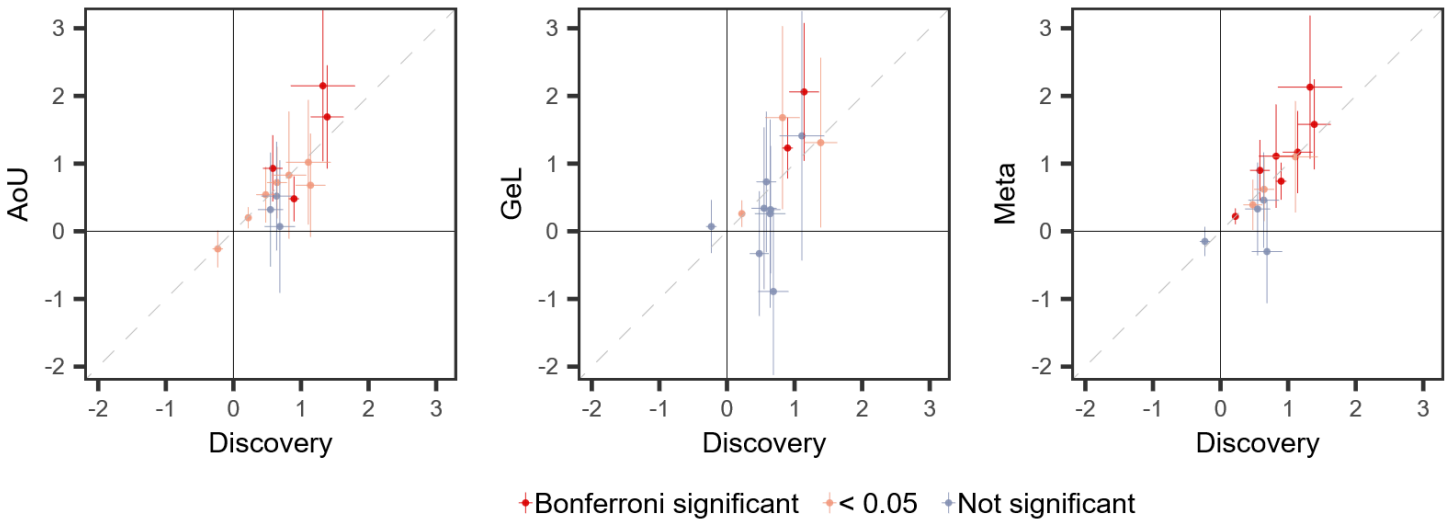

**Supplementary Figure 3: Quantile-quantile plots of the main discovery analysis for mask-based burdens and the gene-based Cauchy combination.** All panels are quantile-quantile plots with expected -log10 *P*-values on the x-axis and observed values on the y-axis. Part **a** shows association results for the main discovery meta-analysis including all ancestries (ALL), while part **b** shows results for a sensitivity analysis restricting only to individuals of genetically-predicted European ancestry (EUR). In both parts, the left panel shows results for mask-based burden testing, with each dot representing a single mask (as such, each gene may be represented up to 60 times), while the right panel shows results for the Cauchy combination approach, and therefore each dot represents a single gene. Lambda values (representing the observed chi-squared statistic over the expected chi-squared statistic at a given distributional cutoff) are added to the plots, including the lambda value at the 90th and 95th percentile of the test statistic distribution.



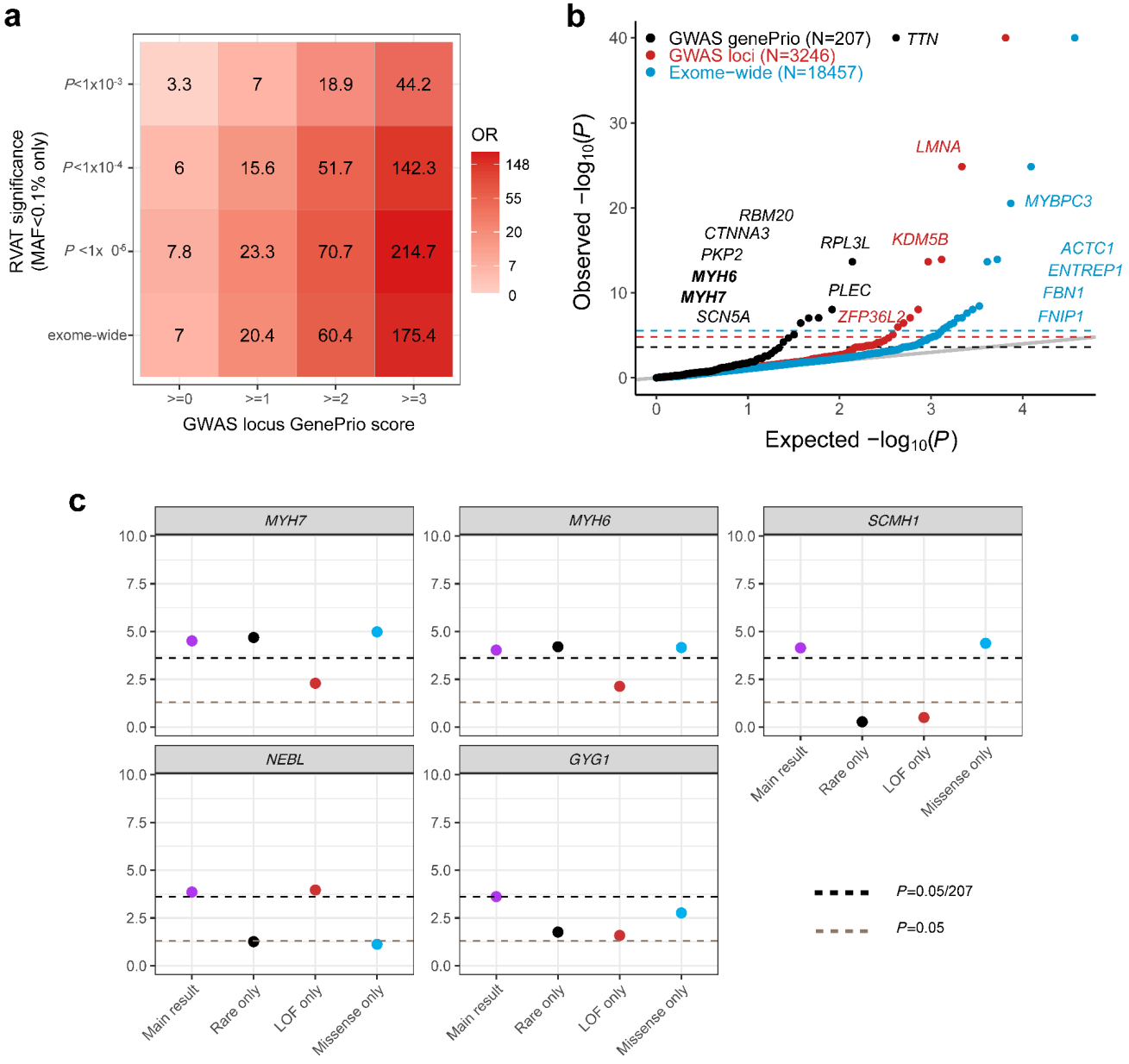

**Supplementary Figure 4: Manhattan plots for sex-stratified meta-analyses.** Both panels are Manhattan plots with chromosome position on the x-axis and -log10 *P*-values on the y-axis. Panel **a** shows results for the Cauchy combination approach (and therefore each dot represents a single gene) for a female-only analysis, while panel **b** shows results for a male-only analysis.




**Supplementary Figure 5: Quantile-quantile plots for sex-stratified meta-analyses.** All panels are quantile-quantile plots with expected -log10 *P*-values on the x-axis and observed values on the y-axis. Panel **a** shows results for the Cauchy combination approach (and therefore each dot represents a single gene) for a female-only analysis, while panel **b** shows results for a male-only analysis. Lambda values (representing the observed chi-squared statistic over the expected chi-squared statistic at a given distributional cutoff) are added to the plots, including the lambda value at the 90th and 95th percentile of the test statistic distribution.

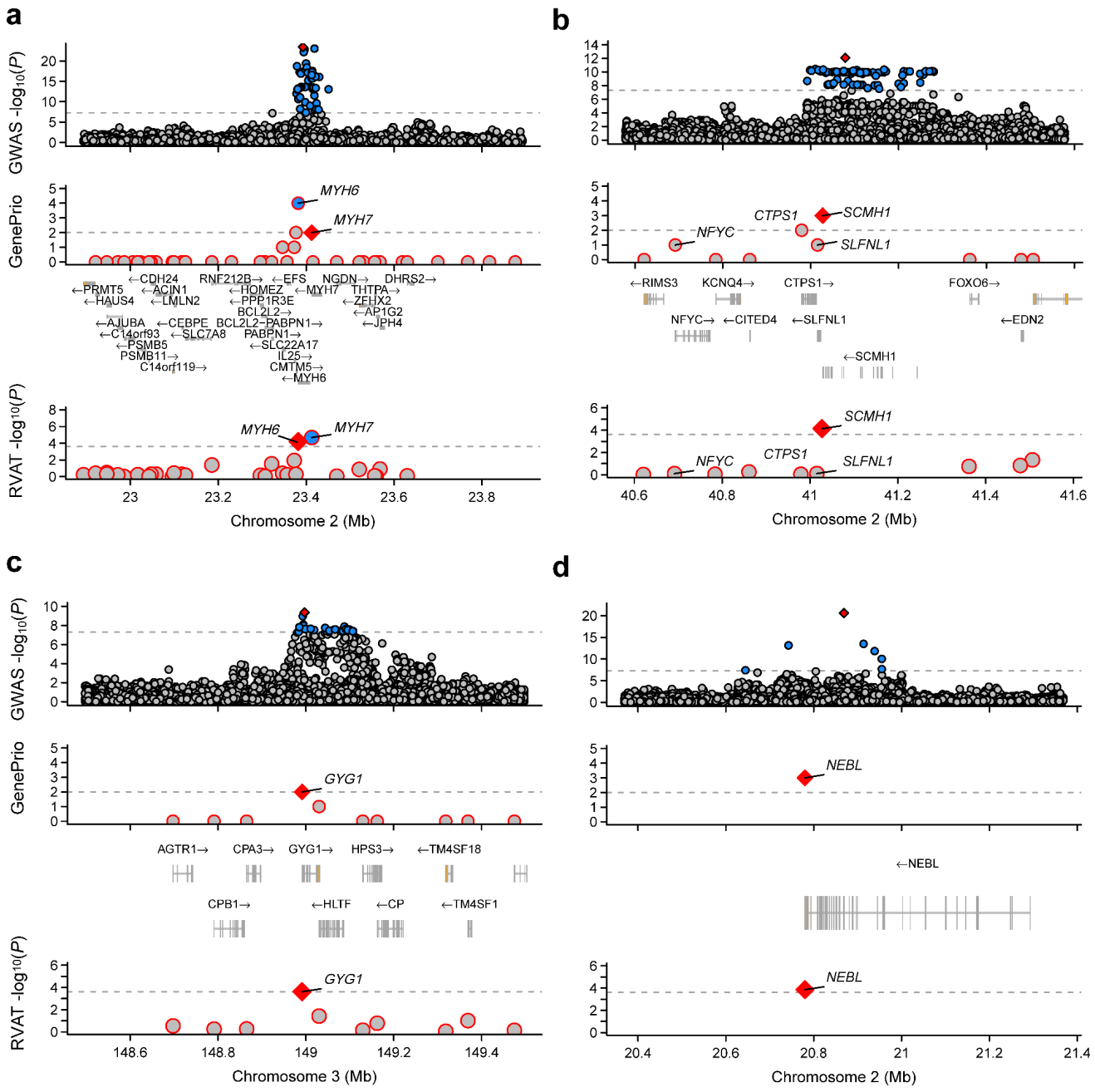



**Supplementary Figure 6: Effect size plot for 15 significant genes across discovery and replication sets.** In all panels, the x-axis shows beta coefficients (with 95% confidence intervals) from the main discovery meta-analysis, while y-axis shows beta coefficients (with 95% confidence intervals) from replication analysis; left panel shows results among separate samples from the All of Us Research Program (AoU; samples from freeze 8, excluding those residing within Massachusetts); middle panel shows results from 100,000 Genomes Project from Genomics England (GeL); right plot shows results from a meta-analysis of these datasets. Each dot represents a single mask from a single gene, where the chosen mask is always the mask that attained the strongest significance in discovery. Red dots were significant in the replication set at two-sided $P<0.05$, while gray dots did not two-sided $P<0.05$ in replication. The dotted gray line represents the y=x line.

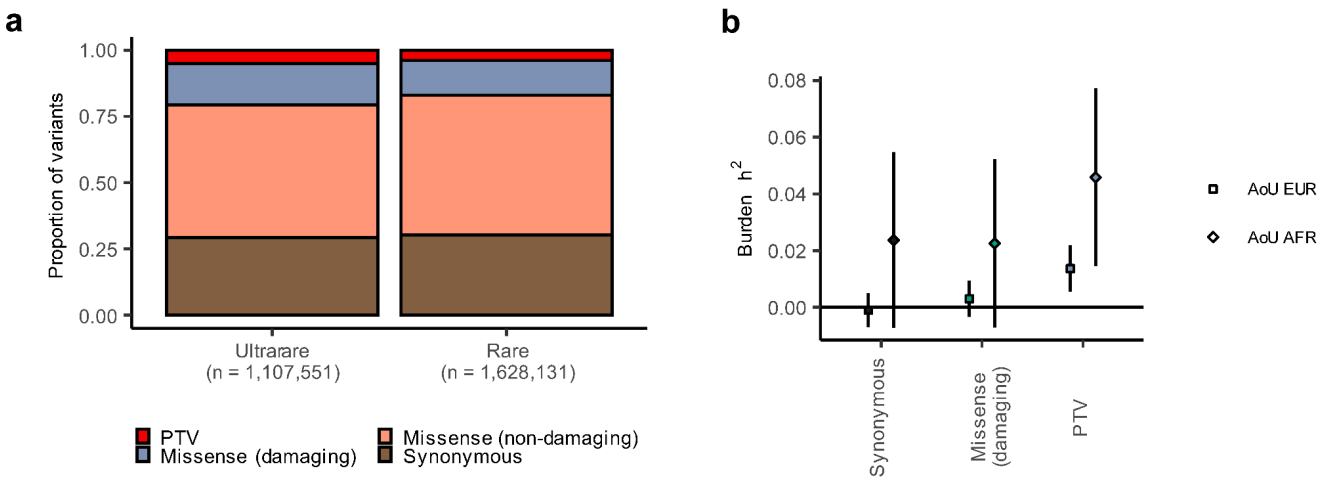



**Supplementary Figure 7: Convergence of RVAS and GWAS using more stringent frequency separation.** Panel **a** is a heatmap showing Fisher exact test enrichments of strict RVAS genes and GWAS genes. The y-axis shows different significance thresholds (based on gene-based Cauchy *P*-values) to determine strict RVAS genes, while the x-axis shows different cutoffs for GenePrio scores for genes near loci from a recent AF GWAS[9] (where GenePrio score of 0 includes any gene within 500kb from genome-wide significant loci). The cell values show ORs from Fisher exact tests. Notably, strict RVAS gene signals here are based only on masks restricting to variants with $MAF_{max}$<0.1%, and therefore does not include potentially low-frequency variants (0.1%>$MAF_{max}$<1%). Panel **b** is a quantile-quantile plot showing results from our strict RVAS analysis (again restricting only to masks with variants with $MAF_{max}$<0.1%), with expected -log10 *P*-values on the x-axis and observed gene-based -log10 Cauchy *P*-values on the y-axis. The black dots represent Cauchy *P*-values for genes residing within GWAS loci that attained GenePrio scores >=2; red dots represent Cauchy *P*-values for all genes residing within GWAS loci; blue dots represent Cauchy *P*-values for all assessed genes exome-wide. Among the black dots, all genes surpassing the Bonferroni-corrected

significance threshold (*P*<0.05/207; shown in black dotted line) are annotated with their gene name, with genes not identified in the exome-wide discovery analysis highlighted in bold; among the red dots, genes surpassing the Bonferroni-corrected threshold (*P*<0.05/3246; red dotted line) are annotated with the gene name if they were not among the genes with high GenePrio score; among the blue dots, genes surpassing the exome-wide Bonferroni-corrected threshold (*P*<0.05/18457; black dotted line) are annotated with gene names if the genes were not near any GWAS loci. Of note, analyses in panels **a** and **b** are restricted to autosomal genes. Part **c** represents a multi-panel dot plot, showing several important sensitivity analyses performed for 5 genes that were significant in an RVAS analysis restricting to genes highly prioritized from GWAS data (genePrio>=2; 207 total genes tested). The y-axis again represents the -log10 of the gene-based *P*-values, while on the x-axis four different analyses are shown with respective Cauchy *P*-values - the original main analysis, an analysis restricting to rare genetic variants (that is, including only masks with $MAF_{max}$<0.1% or $MAF_{max}$<0.001%), an analysis restricting only to PTV variant masks, and an analysis restricting only to missense variant masks. The black dotted line represents the Bonferroni significance cutoff for the specific approach (0.05/207), while the gray dotted line represents the nominal significance cutoff (*P*<0.05). Abbreviations: RVAT, rare variant association testing; GWAS, genome-wide association study; LOF, loss-of-function variants (same as protein-truncating variants).




**Supplementary Figure 8: Regional plots and prioritization across GWAS loci showing subthreshold RVAS signals.** All panels show LocusZoom[62] plots, where the top track shows single variants with -log10 *P*-values on the y-axis and genomic position on the x-axis (with colour indicating LD with respect to the sentinel variant in the locus), the middle track shows genes in the locus with GenePrio scores on the y-axis and genomic position on the x-axis, and the bottom track shows genes in the locus with RVAS Cauchy *P*-values (from the discovery analysis) on the y-axis and position on the x-axis. Panel **a** shows the *MYH6*/*MYH7* locus, panel **b** shows the *SCMH1* locus, panel **c** shows the *GYG1* locus, and panel **d** shows the *NEBL* locus. Abbreviations: RVAT, rare variant association testing; GWAS, genome-wide association study; Mb, megabases.




**Supplementary Figure 9: Burden score regression results from the All of Us Research Program.** Panel **a** is a stacked bar chart showing the proportions of different annotation classes among the variants included in BHR in the All of Us analyses, with the left bar showing ultra-rare variants and the right bar showing rare variants. Red indicates PTVs, blue indicates predicted-damaging missense variants, peach indicates predicted non-damaging missense variants and brown indicates synonymous variants. Panel **b** is a dot plot showing results from the BHR analysis in All of Us, with estimated liability-scale $h^2_{burden}$ for AF on the y-axis, and different masks (split by annotation and frequency) on the x-axis. Squares show results for European genetic ancestry, while the diamonds show results for African genetic ancestry. Error bars represent 95% confidence intervals. The colour scheme is identical to panel a. Abbreviations: PTV, protein-truncating variants; EUR, European genetic ancestry; AFR, African genetic ancestry; AoU, All of Us Research Program.

## Supplementary Tables

*Supplementary tables are presented in the excel sheets.*

# References


1. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
2. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
3. Backman, J. D. *et al.* Exome sequencing and analysis of 454,787 UK Biobank participants. *Nature* **599**, 628–634 (2021).
4. Li, S., Carss, K. J., Halldorsson, B. V., Cortes, A. & UK Biobank Whole-Genome Sequencing Consortium. Whole-genome sequencing of half-a-million UK Biobank participants. *medRxiv* 2023.12.06.23299426 (2023) doi:10.1101/2023.12.06.23299426.
5. Jurgens, S. J. *et al.* Analysis of rare genetic variation underlying cardiometabolic diseases and traits among 200,000 individuals in the UK Biobank. *Nat Genet* **54**, 240–250 (2022).
6. Dewey, F. E. *et al.* Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science* **354**, (2016).
7. Carey, D. J. *et al.* The Geisinger MyCode community health initiative: an electronic health record-linked biobank for precision medicine research. *Genet Med* **18**, 906–913 (2016).
8. Cronin, R. M. *et al.* Development of the Initial Surveys for the All of Us Research Program. *Epidemiology* **30**, 597–608 (2019).
9. Ramirez, A. H. *et al.* The Research Program: Data quality, utility, and diversity. *Patterns (N Y)* **3**, 100570 (2022).
10. Ruff, C. T. *et al.* Evaluation of the novel factor Xa inhibitor edoxaban compared with warfarin in patients with atrial fibrillation: design and rationale for the Effective aNticoaGulation with factor xA next GEneration in Atrial Fibrillation-Thrombolysis In Myocardial Infarction study 48 (ENGAGE AF-TIMI 48). *Am Heart J* **160**, 635–641 (2010).
11. Giugliano, R. P. *et al.* Edoxaban versus warfarin in patients with atrial fibrillation. *N Engl J Med* **369**, 2093–2104 (2013).
12. Scirica, B. M. *et al.* Saxagliptin and cardiovascular outcomes in patients with type 2 diabetes mellitus. *N*

*Engl J Med* **369**, 1317–1326 (2013).

13. Scirica, B. M. *et al.* The design and rationale of the saxagliptin assessment of vascular outcomes recorded in patients with diabetes mellitus-thrombolysis in myocardial infarction (SAVOR-TIMI) 53 study. *Am Heart J* **162**, 818–825.e6 (2011).
14. Bonaca, M. P. *et al.* Long-term use of ticagrelor in patients with prior myocardial infarction. *N Engl J Med* **372**, 1791–1800 (2015).
15. Bonaca, M. P. *et al.* Design and rationale for the Prevention of Cardiovascular Events in Patients With Prior Heart Attack Using Ticagrelor Compared to Placebo on a Background of Aspirin-Thrombolysis in Myocardial Infarction 54 (PEGASUS-TIMI 54) trial. *Am Heart J* **167**, 437–444.e5 (2014).
16. Sabatine, M. S. *et al.* Rationale and design of the Further cardiovascular OUtcomes Research with PCSK9 Inhibition in subjects with Elevated Risk trial. *Am Heart J* **173**, 94–101 (2016).
17. Sabatine, M. S. *et al.* Evolocumab and Clinical Outcomes in Patients with Cardiovascular Disease. *N Engl J Med* **376**, 1713–1722 (2017).
18. Wiviott, S. D. *et al.* Dapagliflozin and Cardiovascular Outcomes in Type 2 Diabetes. *N Engl J Med* **380**, 347–357 (2019).
19. Taliun, D. *et al. Sequencing of 53,831 Diverse Genomes from the NHLBI TOPMed Program*. (2021).
20. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
21. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655–1664 (2009).
22. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
23. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
24. Conomos, M. P., Miller, M. B. & Thornton, T. A. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet Epidemiol* **39**, 276–293 (2015).

25. Conomos, M. P., Reiner, A. P., Weir, B. S. & Thornton, T. A. Model-free Estimation of Recent Genetic Relatedness. *Am J Hum Genet* **98**, 127–148 (2016).
26. Akbari, P. *et al.* Multiancestry exome sequencing reveals INHBE mutations associated with favorable fat distribution and protection from diabetes. *Nat Commun* **13**, 4844 (2022).
27. Van Hout, C. V. *et al.* Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, 749–756 (2020).
28. Van der Auwera, G. A. & O'Connor, B. D. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. (O'Reilly Media, 2020).
29. Gómez Vela, F. A., Divina, F. & García-Torres, M. *Computational Methods for the Analysis of Genomic Data and Biological Processes*. (MDPI, 2021).
30. Galinsky, K. J. *et al.* Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. *Am J Hum Genet* **98**, 456–472 (2016).
31. Jurgens, S. J. *et al.* Rare coding variant analysis for human diseases across biobanks and ancestries. *Nat Genet* **56**, 1811–1820 (2024).
32. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122 (2016).
33. Liu, X., Li, C., Mou, C., Dong, Y. & Tu, Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Medicine* **12**, 1–8 (2020).
34. Ioannidis, N. M. *et al.* REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet* **99**, 877–885 (2016).
35. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
36. Gao, H. *et al.* The landscape of tolerated genetic variation in humans and primates. *Science* **380**, eabn8153 (2023).
37. Orenbuch, R. *et al.* Deep generative modeling of the human proteome reveals over a hundred novel genes involved in rare genetic disorders. *medRxiv* (2023) doi:10.1101/2023.11.27.23299062.