SUPPLEMENTAL MATERIAL

Table of Contents

Supplementary Methods

Summary of study populations and gene sequencing

We analyzed gene sequencing data from four case-control study variant call sets, Supplemental Table I-V. Inclusion of both cases and controls from a given study in joint calling of DNA variants minimizes potential sequencing artifacts or batch effects.

First, we performed whole-exome sequencing in 24,097 cases and 30,354 controls from the Myocardial Infarction Genetics ExSeq study ('MIGen ExSeq'), which involves participants from 11 case-control studies, Supplemental Table I, II and Supplemental Figure IA.

Second, we performed whole-genome sequencing in 2,369 cases and 4,218 controls from the Myocardial Infarction Genetics WGSeq study ('MIGen WGSeq') after sample quality control, which includes participants from two case-control studies – one that was derived from a combination of the VIRGO (Variation in Recovery: Role of Gender on Outcomes of Young AMI Patients) and Multi-Ethnic Study of Atherosclerosis (MESA) studies as previously described and the second from the Taiwanese TAICHI study[5,45] (Supplemental Table III and Supplemental Figure IB).

Third, we performed whole-exome sequencing of 6,446 cases and 5,932 control participants after sample quality control from the UK Biobank prospective cohort study ('UK Biobank 13K')[16,17], (Supplemental Table IV and Supplemental Figure IC).

Fourth, we downloaded whole-exome sequencing data from the 8,169 cases and 176,611 controls after sample quality control from the UK Biobank study generated by the Regeneron Genetics Center ('UK Biobank 200K')[18], (Supplemental Table V and Supplemental Figure ID). Individuals included in the UK Biobank 13K analysis were removed from the UK Biobank 200K study to ensure no participant overlap.

Coronary artery disease cases referred to individuals who suffered myocardial infarction, underwent coronary revascularization, had angiographically confirmed obstructive disease, or died from coronary causes. Additional details of case ascertainment strategies and study populations are provided in details as below.

Summary of quality control

Within each of the four call sets, participant samples were removed from analysis based on excessive DNA contamination, inadequate sequencing coverage, sample duplicates, low variant call rate, discordance between reported- and genotype-based sex assessments, or relatedness (second-degree relative or closer) to another study participant. These parameters led to the exclusion of 19,343 participants across the four call sets, which includes 2,727 (4.8%) of 57,178 individuals from the MIGen ExSeq study, 222 (3.3%) of 6,809 participants from the MIGen WGSeq study, 531 (4.1%) of

12,909 from the UK Biobank 13K study, and 15,863 (7.9%) of 200,643 from the UK Biobank 200K study, respectively. In the UK Biobank 200K study, 14,566 (7.2%) of 200,643 were removed due to the sample relatedness, details in below section. After sample quality control, there are 41,081 cases and 217,115 controls available for analysis. Of the 41,081 cases, only 5,791 (14%) were included in our previous rare variant association study across all protein-coding genes[4].

After completion of sample-level quality control, variants were removed based on failures of a previously described Variant Quality Score Recalibration algorithm[46], presence in regions of the genome that limit accurate read alignment, low variant quality by depth score, low call rate, differential missingness of a variant between cases and controls of a given study, and ancestry-specific Hardy Weinberg disequilibrium as performed previously[16,46], details see below section.

Summary of rare variant annotation and aggregation strategies

We first aggregated predicted loss-of-function mutations within a given gene based on two annotations: (i) ultra-rare (allele frequency <0.01%) variants annotated as high-confidence inactivating variants – stop-gained, splice-disrupting, or frameshift – by the Loss-Of-Function Transcript Effect Estimator (LOFTEE) algorithm or a cryptic splice site predicted by the SpliceAI algorithm[39,47]; (ii) variants annotated as pathogenic or likely pathogenic in the ClinVar online database with no conflicting reports suggesting benign or uncertain significance and maximum population allele frequency <0.1% in the Genome Aggregation Database (gnomAD) [47,48]. High-confidence inactivating and ClinVar variants were assigned a score of 1 – corresponding to predicted complete inactivation of this allele – and Splice AI variants a score of 0.75 – corresponding to predicted 75% inactivation based on previous recommendations[39]. This aggregation strategy will be referred to hereafter as putative loss-of-function ('pLoF').

To increase statistical power, we next integrated ultra-rare (allele frequency <0.01%) missense variants predicted to be damaging by each of five computational algorithms – SIFT, Polyphen2-HDIV, Polyphen2-HVAR, LRT, and MutationTaster as performed previously[4,8]. Given that these variants are known to be less damaging to protein function than loss-of-function variants, missense variants were assigned a gene-specific weight that took into account the cumulative frequencies of high-confidence loss-of-function variants as compared to that of predicted damaging missense variants as previously described[49,50], details see the below section. This aggregation strategy will be referred to hereafter as putative loss-of-function plus missense ('pLoF+missense').

Summary of statistical analysis

We tested the association between the aggregated rare variant burden score of each gene and risk of disease using Firth logistic regression, a test robust to association testing in the context of low carrier counts or case-control imbalance[51], implemented in the R “SPAtest” package. An individual’s gene-specific score was computed as the weighted sum of the total number of qualified rare variants according to the weighting strategy described above and capped at one. All analyses included genetic ancestry – as quantified by the first 10 principal components – as covariates in the regression analysis[52], with additional inclusion of individual cohort indicator in the MIGen ExSeq study and age and sex (except when uses as part of ascertainment scheme, Supplemental Table X). The results for each gene were combined across the four datasets using fixed-effects meta-analysis with effects estimated from the Firth logistic regression, with absence of significant heterogeneity (p-heterogeneity > 0.05) confirmed across datasets for all significantly associated genes reported in this study. The meta-analysis and heterogeneity test was performed by the Metasoft software[53]. A sensitivity analysis for the *NOS3* gene was performed by partitioning the samples in each data set into two groups according to European or non-European ancestry, given that each dataset was dominated by only European or one non-European ancestry (Supplemental Tables I-V). The association was then tested in each group using the same variant as in the main analysis. Finally, the results were summarized by a fixed-effect meta-analysis. A gene was recognized as genome-wide significant if the gene’s *P-value* was smaller than the Bonferroni-corrected p-value threshold of 1.25×10^{-6}, assuming 20,000 genes tested and two variant masks used in this study. All P-Value reported in this study are two-sided and without multiple test corrections.

Within the UK Biobank datasets, biomarker concentrations for carriers versus noncarriers of a given gene were compared using concentrations standardized to mean for covariates of age, sex, and top 10 principal components of ancestry, Supplemental Table VI and Supplemental Table X. The same covariate structure was used to compare the rate of hypertension between carriers with noncarriers. The estimated untreated lipids and blood pressure levels were used for all of the analyses. For the MIGen ExSeq and the MIGen WGSeq studies, the estimated untreated LDL cholesterol and Triglycerides were estimated by dividing measured value by 0.70 and 0.85, respectively, if taking lipid-lowering medicine assuming statin[54]. For the UK Biobank 13K and the UK Biobank 200K studies, the adjustments were made according to the type of lipid-lowering medication intake, detail see *eTable 1* of Patel, A. P. et al.’s study[54]. For the blood pressure, the estimated untreated blood pressure levels were estimated by adding 15 mm Hg to systolic and 10 mm Hg to diastolic values, respectively, to participants who reported use of blood-pressure lowering medications[55,56].

All the association tests and regression analyses were performed in R, version 3.6.1. Details for the data availability and code availability please see the below section.

Study populations, gene sequencing, and quality control

We analyzed gene sequencing data from four case-control study variant call sets: (i) the Myocardial Infarction Genetics ExSeq study; (ii) the Myocardial Infarction Genetics WGSeq study; (iii) the UK Biobank 13K study; and (iv) the UK Biobank 200K study.

Myocardial Infarction Genetics ExSeq study

The Myocardial Infarction Genetics ExSeq (MIGen ExSeq) study aggregated whole-exome sequencing data from 57,178 individuals – 24,678 cases affected by coronary artery disease and 32,500 controls – derived from 11 case-control studies (Supplemental Table I and II). Cases were individuals affected by coronary artery disease, defined as myocardial infarction, coronary revascularization procedure, or angiographically confirmed coronary artery disease as previously described[43,57–68]. To select for early-onset cases, a subset of the studies included age-of-onset as an additional criterion (Supplemental Table II).

Whole-exome sequencing was performed at the Broad Institute of MIT and Harvard (Cambridge, MA, USA) for all studies – except for the Atherosclerosis Risk in Communities Study where sequencing was performed at the Baylor College of Medicine Human Genome Sequencing Center – as previously described[43,60,69]. In brief, sequence data of all participants were aligned to the human reference genome build GRCh37.p13 using the Burrows-Wheeler Aligner algorithm[70]. Aligned non-duplicate reads were locally realigned, and base qualities were recalibrated using Genome Analysis Toolkit (GATK) software[71]. Variants were jointly called using GATK HaplotypeCaller module[72]. The individual genotype call was set as missing if reads depth (DP) ≤ 10 or DP ≥ 200, if homozygous reference allele with genotype quality (GQ) ≤ 20 or the ratio of alt allele reads over all covered reads > 0.1, if heterozygous with the ratio of alt allele reads over all covered reads < 0.2 or Phred-scaled likelihood (PL) of the reference allele < 20, or if homozygous alternate with the ratio of alt allele reads over all covered reads < 0.9 or PL of reference allele < 20.

Of the 57,178 individuals who underwent sequencing, 2,727 (4.8%) were removed based on sequencing quality control metrics. Sample exclusion criteria included:

- DNA Contamination > 10% (N = 12 samples removed)
- Genotype/phenotype sex discordance or ambiguous sex, definied as 0.5 < Fstat[73] < 0.8 (N = 525 samples removed)
- Ancestry specific excess heterozygosity, as defined by F coefficient in PLINK[73] (N = 32 samples removed)
- Mean sequencing depth (DP)< 30x (N = 14 samples removed)
- Average genotyping quality (GQ), mean GQ < 80 (N = 78 samples removed)
- Genotype call rate < 90% (N = 97 samples removed)

- Second-degree relative or closer with another study participant, defined as kinship coefficient > 0.0884[74] (N = 1,969 samples removed).

After completion of sample level quality control, variant quality control was performed using the following exclusion criteria:

- Failure by the GATK Variant Quality Score Recalibration (VQSR) metric[71], a machine learning algorithm designed to balance sensitivity (calling genuine variants) and specificity (limit false positive variant calls). Default settings were used – corresponding to a truth sensitivity of 99.6% for single nucleotide polymorphisms (SNPs) and 95% for insertion-deletion (indel) variants – except for single nucleotide polymorphisms present in only a single individual ('singleton'). Because singletons are known to have a deflated VQSR score, we were slightly more permissive and additionally included variants with a truth sensitivity of 99.8%.
- Variants in low-complexity regions of the genome that preclude accurate read alignment as previously defined[75].
- Variants in segmental duplications of the genome[75,76].
- Quality by depth score < 2 (for single nucleotide polymorphisms) or < 3 (for insertion-deletions)
- Ancestry-specific Hardy-Weinberg disequilibrium p-value < $1x10^{-8}$.
- Variants were set to missing within specific cohort/exome-capture combinations if the per cohort/exome-capture call rate was < 90% or missingness was non-random between cases and controls per combination as tested with a chi-square test[73] (P-value < $5x10^{-9}$) or per combination variant call rate < 90%.

Following the application of these exclusion criteria, 8,716,575 high-quality variants were carried forward into the analysis.

Myocardial Infarction Genetics WGSeq study

Myocardial Infarction Genetics WGSeq (MIGen WGSeq) study performed whole-genome sequencing in 6,809 participants with samples from two case-control studies – one previously described derived from the VIRGO (Variation in Recovery: Role of Gender on Outcomes of Young AMI Patients) and Multi-Ethnic Study of Atherosclerosis (MESA) studies and the second from the Taiwanese TAICHI study[5,45].

The design and whole-genome sequencing of the VIRGO-MESA case-control study has been previously described[5,77,78]. In brief, we aggregated whole-genome sequencing data from 2,101 coronary artery disease cases derived from the VIRGO study and 3,932 controls derived from the MESA study. The VIRGO study enrolled participants hospitalized with acute myocardial infarction, aged 18 to 55 years, who were enrolled between 2009 and 2012 from 103 United States and 24

Spanish hospitals using a 2:1 female-to-male enrollment design. Baseline patient data were collected by medical chart abstraction and standardized in-person patient interviews administered by trained personnel during the index acute myocardial infarction admission. Individuals with available DNA and who had provided written informed consent for genetic analysis were included in the present study. The design of the MESA study has been previously described and protocol available at www.mesa-nhlbi.org. In brief, participants between the ages of 45 and 84 without prevalent cardiovascular disease were recruited between 2000-2002 from 6 United States communities. Individuals were excluded if informed consent for genetic testing had not been obtained/was withdrawn, DNA was not available for sequencing, or incident cardiovascular disease (myocardial infarction, coronary revascularization, angina, peripheral arterial disease, stroke, resuscitated cardiac arrest, death due to cardiovascular causes) through the period of last available follow-up in December 2014.

We additionally successfully performed whole-genome sequencing in 288 coronary artery disease cases and 457 controls derived from the TAICHI study. The TAICHI case-control study recruited Taiwanese Chinese individuals at four academic centers[45]. Cases with coronary disease were identified as those with a history of myocardial infarction, coronary revascularization, or stenosis of ≥ 50% in a major epicardial vessel demonstrated by angiography. All cases experienced an early-onset coronary event (men ≤ 50 years, women ≤ 60 years) in the context of normal circulating lipid levels (LDL cholesterol < 130 mg/dl or total cholesterol < 185 mg/dl). Control subjects with no prior history of CAD were enrolled from an epidemiology study and from the several Hospital Endocrinology and Metabolism Departments either as outpatients or as their family members. Subjects with a history of CAD were excluded.

Whole-genome sequencing was performed at the Broad Institute of MIT and Harvard(Cambridge, MA, USA) as previously described[5]. In brief, libraries were constructed and sequenced on the Illumina HiSeqX with the use of 151-bp paired-end reads for whole-genome sequencing. Output from Illumina software was processed by the Picard data-processing pipeline(http://broadinstitute.github.io/picard/) to yield BAM files containing well-calibrated, aligned reads. A sample was considered sequence complete when the mean coverage was ≥ 30x (for the MESA cohort) or ≥ 20x (for the VIRGO and TAICHI cohort). Sample genotypes were determined via a joint callset using the GATK HaplotypeCaller module. The individual genotype call was set as missing if reads depth (DP) ≤ 10 or DP ≥ 200, if homozygous reference allele with genotype quality (GQ) ≤ 20 or the ratio of alt allele reads over all covered reads > 0.1, if heterozygous with the ratio of alt allele reads over all covered reads < 0.2 or Phred-scaled likelihood (PL) of the reference allele < 20, or if homozygous alternate with the ratio of alt allele reads over all covered reads < 0.9 or PL of reference allele < 20.

Analysis of a joint call set for the VIRGO-MESA study – where sequencing was performed in separate batches – was performed to minimize any potential confounding[5]. As described previously, we confirmed that the overall number of variants per individual was similar between VIRGO cases and MESA control subjects in ancestry-stratified analysis and performed an association study of all observed common (allele frequency ≥ 1%) variants, confirming no significant inflation of test statistics[5].

Of the 6,809 individuals who underwent whole-genome sequencing, 222 (3.3%) were excluded based on sequencing quality control metrics. Sample exclusion criteria included:

- Duplicate removal (n = 15 samples removed)
- DNA Contamination > 5% (n = 21 samples removed)
- Mean DP < 20x (n = 3 samples removed)
- Genotype/phenotype sex discordance or ambiguous sex ($0.5 < F_{stat} < 0.8$, n = 7 samples removed).
- Call Rate < 95% (n = 24 samples removed)
- Removed one from each pair of related (defined by second degree of relationship or closer[74], n = 152 samples removed)

After completion of sample level quality control, variant quality control was performed using the following exclusion criteria:

- Failure by the GATK VQSR metric. Default settings were used – corresponding to a truth sensitivity of 99.8% for single nucleotide polymorphisms (SNPs) and 99.95% for insertion-deletion (indel) variants.
- Variants in low-complexity regions[75].
- Variants in segmental duplication region of the genome[75,76].
- Quality by depth score < 2 (for SNPs) or < 3 (for indels).
- Variant call rate < 95%.
- Ancestry-specific Hardy-Weinberg disequilibrium p-value < $1x10^{-6}$.

Following the application of these exclusion criteria, 145,897,548 high-quality variants were carried forward into the analysis.

UK Biobank 13K Study

The UK Biobank 13K study involved whole-exome sequencing data from 12,909 participants – 6,472 coronary artery disease cases and 6,437 controls – as previously described[16]. Participants were derived from the UK Biobank, a prospective national biobank study that enrolled middle-aged adult participants between 2006 and 2010. Coronary artery disease cases were defined centrally based on self-report at enrollment, hospitalization records, or death registry records (http://Biobank.ndph.ox.ac.uk/showcase/showcase/docs/ alg_outcome_mi.pdf). Controls included participants free of any self-reported or documented history of coronary artery disease.

Circulating lipid biomarker concentrations were measured from blood samples taken at time of enrollment as part of the study protocol. For participants who reported use of lipid-lowering medications, lipid levels were adjusted depending on the type of lipid-lowering medication intake based on prior reports of effect size for each medication type from the literature as reported previously[54], eTable 1 in Patel AP *et al.*[54]. For example, in the case of statin intake, total cholesterol was divided by 0.8, lower density cholesterol by 0.7, and triglycerides by 0.85. Systolic blood pressure was measured in study participants at the time of enrollment. For participants who reported the use of blood-pressure lowering medications, we estimated untreated values by adding 15 mm Hg to systolic and 10 mm Hg to diastolic values as performed previously[55,56].

Whole-exome sequencing was performed at the Broad Institute of MIT and Harvard (Cambridge, MA) as described previously[16]. In brief, libraries were constructed and sequenced on Illumina HiSeq sequencing using 151 bp pair-end reads. An Illumina Nextera Exome Kit was used for in-solution hybrid selection. Sequencing reads were aligned to the human reference genome build GRCh37.p13 using the Burrows–Wheeler Aligner algorithm, and aligned non-duplicate reads were locally realigned and base quantiles were recalibrated using the GATK software[70,71]. The individual genotype call was set as missing if reads depth (DP) ≤ 10 or DP ≥ 200, if homozygous reference allele with genotype quality (GQ) ≤ 20 or the ratio of alt allele reads over all covered reads > 0.1, if heterozygous with the ratio of alt allele reads over all covered reads < 0.2 or Phred-scaled likelihood (PL) of the reference allele < 20, or if homozygous alternate with the ratio of alt allele reads over all covered reads < 0.9 or PL of reference allele < 20.

Of the 12,909 individuals who underwent sequencing, 531 (4.1%) were removed based on sequencing quality control metrics. Sample exclusion criteria included:

- DNA Contamination > 10% (N = 0 samples removed)
- Genotype/phenotype sex discordance or ambiguous sex, defined as 0.5 < Fstat < 0.8[73] (N = 17 samples removed)
- Excess heterozygosity, as defined by F coefficient in PLINK[73] (N = 4 samples removed)
- Genotype call rate < 95% (N = 6 samples removed)
- Mean sequencing depth < 30x (N = 0 samples removed)
- Average genotyping quality < 80 (N = 0 samples removed)
- Second-degree relative or closer with another study participant, defined as kinship coefficient > 0.0884[74] (N = 503 samples removed)
- Withdrawal of informed consent (N = 1 samples removed)

After completion of sample level quality control, variant quality control was performed using the following exclusion criteria:

- Failure by the GATK VQSR metric. Default settings were used – corresponding to a truth sensitivity of 99.6% for single nucleotide polymorphisms (SNPs) and 95% for insertion-deletion (indel) variants – except for single nucleotide polymorphisms present in only a single individual ('singleton'). Because singletons are known to have a deflated VQSR score, we were slightly more permissive and additionally included variants with a truth sensitivity of 99.8%.
- Variants in low-complexity regions of the genome that preclude accurate read alignment as previously defined[75].
- Variants in segmental duplication region of the genome[75,76].
- Quality by depth score < 2 (for single nucleotide polymorphisms) or < 3 (for insertion-deletions)
- Ancestry-specific Hardy-Weinberg disequilibrium p-value < $1x10^{-6}$.
- Variant call rate < 95%.

UK Biobank 200K study

The UK Biobank 200K study involved analysis of whole-exome sequencing data from an additional 200,643 participants. Coronary artery disease was defined based on self-report of heart attack/myocardial infarction, hospitalization records confirming a diagnosis of acute myocardial infarction or ischemic heart disease coronary revascularization procedures (coronary artery bypass graft surgery or percutaneous angioplasty/stent placement), or death registry data indicating ischemic heart disease or myocardial infarction as a cause of death as previously described [79].

Circulating lipid biomarkers and blood pressures were measured, and levels were adjusted the same way as above for the UK Biobank 13K participants.

Whole-exome sequencing was performed by the Regeneron Genetics Center using an updated Functional Equivalence (FE) protocol that retains original quality scores in the CRAM files (referred to as the OQFE protocol) as previously described[18]. The DTxGen Exome Research Panel v1.0 including supplemental probes was used for exome capture for this data set (https://biobank.ctsu.ox.ac.uk/showcase/label.cgi?id=170). 19,396 genes in the targets of 38Mbp were covered. 75x75bp paired-end reads were sequenced on the Illumina NovaSeq 6000 platform. For each sample in the targeted region, more than 95.2% of sites were covered by more than 20 reads. The call set was delivered in the PLINK file format. We converted the file to the VCF format by the plink2 software (version: v2.00a3LM 64-bit Intel)[73], and lifted it over to GRCh37 using CrossMap software (version: v0.3.3)[80]. As for the above datasets, we filtered the variants in the low-complexity regions, in the segmental duplication regions, or in the decoy regions[75,76]. Further, 184 samples were excluded due to lack of the principal component of ancestry values from UK Biobank central QC[81]. 1,113 samples were excluded due to overlap with samples in the UK Biobank 13K Study. 14, 566 samples were removed due to relatedness (second-degree relative or closer, same procedure as the UK Biobank 13K described above). 184,780 samples were carried forward for the final analysis, including 8,169 cases and 176,611 controls.

Principal component analysis

For the Myocardial Infarction Genetics ExSeq study and Myocardial Infarction Genetics WGSeq study, a panel of approximately 16,000 ancestry informative markers[82] (AIMs) identified across six continental populations[83] was chosen to derive principal components (PCs) of ancestry for all samples that passed quality control. For ancestry inference, samples from the 1000 Genomes project[84] were mixed together for running the principal component analysis (PCA). However, the top PCs used as covariates in the regression analysis were from QC passed samples only and without including samples from the 1000 Genomes project. The principal component analysis (PCA) was performed using EIGENSTRAT[52]. For the UK Biobank samples, we used the PC values derived from UK biobank central QC[81].

Ancestry inference

Within the Myocardial Infarction Genetics Consortium ExSeq and WGSeq studies, a small proportion of participants had no self-reported race or had discordance between self-reported race and genetic ancestry as quantified by principal components. As some QC were ancestry-specific, in order to assign ancestry – African, European, East Asian, South Asian, or Others – to individuals without clear reported ancestry, a k-nearest neighbors (k-NN) classifier was applied using k-NN algorithm implementation from the Scikit-learn library in Python.

For the MIGen ExSeq data, principal components for ancestry inference were derived in 2,470 unrelated samples from the 1000 Genomes project[84] and 55,362 unrelated samples from the MIGen ExSeq study together. Related samples were then projected to the PC space. The k-NN (k = 5) classifier was trained by ⅔ samples from the 1000 Genomes project using the top six PCs to classify the five super population labels defined in the 1000 Genomes project[84]. The remaining ⅓ independent samples were used as a testing data set. The model had 99.6% classification accuracy in the independent testing samples. The model was then applied to all of the samples in the MIGen ExSeq study to infer ancestry.

For the MIGen WGSeq data, the classifier was built using MESA samples after removing 25 individuals with discordant self-reported race and PC ancestry as determined by visual inspection of PC1 and PC2. The remaining MESA samples were split into a training set (n=2,490) and test set (n=1,246). A k-NN (k=5) classifier was built using self-reported race as the dependent variable (1: White/Caucasian, 2: Chinese American, 3: Black/African-American, 4: Hispanic), and PC1 to PC5 as features. The classifier had a 98.1% reclassification rate in the test set, with misclassifications generally occurring for Hispanic individuals. This classifier was then applied to all 6,587 samples to generate inferred race. Inferred race and self-reported race were concordant in 6,383 of 6,576 (97%) of samples with non-missing self-reported race.

For the UK Biobank 13K and 200K studies, the sample ancestry was determined via self-report[17,81].

Approach to variant annotation and weighting

Across each of the four datasets, we applied a uniform annotation strategy for all variants that passed quality control, variants were annotated into four classes: (i) high-confidence inactivating variants; (ii) cryptic splice sites; (iii) ClinVar pathogenic or likely pathogenic variants; (iv) predicted damaging missense variants.

To identify ultra-rare (minor allele frequency < 0.01%) high-confidence predicted inactivating variants, we applied the previously validated Loss-Of-Function Transcript Effect Estimator (LOFTEE) algorithm implemented within the Ensembl Variant Effect Predictor(VEP) software program as a plugin, VEP version 96.0[47,85]. The LOFTEE algorithm identifies stop-gain, splice-site disrupting, and frameshift variants. The algorithm includes a series of flags for each variant class that collectively represent 'low-confidence' inactivating variants. Here, we study only variants that were 'high-confidence' inactivating variants. Within the association testing framework, this class of variants was given a weight of 1.0.

To identify ultra-rare (minor allele frequency < 0.01%) cryptic splice sites, we applied the previously validated SpliceAI algorithm using default parameters[39]. The algorithm is a deep residual neural network that uses the sequence of pre-mRNA transcripts to predict whether each position is a splice donor, splice acceptor, or neither. For each variant, the SpliceAI estimates four delta scores, DS_AG (acceptor gain delta score), DS_AL (acceptor loss delta score), DS_DG (donor gain delta score), and DS_DL (donor loss delta score). Variants with a maximum of these four scores – which range from 0 to 1 – greater or equal to 0.5 was designated as a cryptic splice site. Within the association testing framework, this class of variants was given a weight of 0.75 as previously recommended[39].

To identify rare variants (minor allele frequency < 0.1% in the Genome Aggregation Database, gnomAD database[47,86]) annotated as pathogenic or likely pathogenic in the publicly available ClinVar database[48], we focused on the clinical significance ('CLINSIG') field[47,48]. The database was downloaded (version 20181202) and variants were included if they included the terms 'pathogenic' and/or 'likely pathogenic' and did not have conflicting pathogenicity assessment in or including annotations of 'benign,' 'likely benign,' or 'uncertain significance.' Within the association testing framework, this class of variants was given a weight of 1.0.

To identify ultra-rare (minor allele frequency < 0.01%) predicted damaging missense variants, we included variants predicted to be damaging by each of five computational prediction algorithms as described previously[8,60,87]. In brief, predictions were retrieved from the dbNSFP database[88], version 2.9.3, with the most severe prediction across multiple transcripts used. We focused on five prediction algorithms: SIFT[89] (including variants annotated as damaging), PolyPhen2-HDIV and PolyPhen2-HVAR[90] (including variants annotated as possibly or probably damaging), LRT[91] (including variants annotated as deleterious), and MutationTaster[92] (including variants annotated as disease-causing-automatic or disease-causing). Within the association testing framework, this class of variants was given a gene-specific weight based on the relative frequency of these predicted damaging missense variants as compared to the frequency of high-confidence predicted inactivating variants identified by LOFTEE algorithm using a previously recommended approach[49,50]: given the cumulative allele frequency of all of the LOFTEE high confidence rare variants of a gene (G) as f_L, the cumulative allele frequency of all of the predicted damaging missense variants as f_M, the weight for the missense variants was estimated as $(\frac{f_L \times (1-f_L)}{f_M \times (1-f_M)})^{0.5}$, and capped at 1.0. For genes without LOFTEE high confidence rare variants, the weight for missense variants is 1.0.

Data availability

The UK Biobank data can be applied through the UK Biobank Access Management System, and the phenotypes derived as part of this manuscript will be returned to the UK Biobank for dissemination to approved investigators. The Whole-genome sequencing data of the Myocardial Infarction Genetics WGSeq study (VIRGO, TaiChi, and MESA study) have been uploaded to the database of Genotypes and Phenotypes (dbGaP) repository under accession numbers phs001259, phs001487, and phs001416, respectively. For the Myocardial Infarction Genetics ExSeq study data set, the data can be applied through dbGaP repository by the following accession numbers: phs000280.v7.p1 (ARIC), phs000814.v1.p1 (ATVB), phs001398.v1.p1 (BRAVE), phs000279.v2.p1 (EOMI), phs001000. v1.p1 (Leicester), phs000990. v1.p1 (North German MI), phs000916.v1.p1 (South German MI), phs000806.v1.p1 (OHS), phs000883.v1.p1 (PROCARDIS), phs000917.v1.p1 (PROMIS), and phs000902.v1.p1(Regicor).

Code availability

- **Genome Analysis Toolkit (GATK) pipeline:**
  https://gatk.broadinstitute.org/hc/en-us

- **Loss-Of-Function Transcript Effect Estimator (LOFTEE):** https://github.com/konradjk/loftee
- **Ensembl Variant Effect Predictor(VEP) software:** https://useast.ensembl.org/info/docs/tools/vep/index.html
- **SpliceAI algorithm (version 1.3):** https://github.com/Illumina/SpliceAI
- **Clinvar database (version 20181202):** https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/
- **gnomAD database, version 2.1.1:** https://gnomad.broadinstitute.org/downloads
- **dbNSFP database, version 2.9.3:** https://sites.google.com/site/jpopgen/dbNSFP
- **R "SPAtest" package (version 3.1.2):** https://cran.rstudio.com/web/packages/SPAtest/index.html
- **Metasoft.jar software (v2.0.0):** http://genetics.cs.ucla.edu/meta/
- **Gene burden scoring:** https://github.com/wavefancy/WallaceBroad/blob/master/python/RareVariantAssociationTest/PgenCountScore4GeneMask/PgenCountScore4GeneMask.py

Supplementary Tables

Table I. Clinical characteristics for the Myocardial Infarction Genetics ExSeq study

VARIABLE	CASE	CONTROL	P-Value
N	24097	30354	
Censor age, mean (SD)	50.9 (10.4)	59.7 (15.8)	<0.001
Sex, Male, n (%)	18980 (80.2)	19870 (66.9)	<0.001
Ancestry, n (%)			
African	405 (1.7)	2682 (8.8)	<0.001
East Asian	3 (0.0)	2 (0.0)	
European	7767 (32.2)	13646 (45.0)	
Other	43 (0.2)	38 (0.1)	
South Asian	15879 (65.9)	13986 (46.1)	
Lipid-lowering therapy, n (%)	5674 (30.5)	723 (3.2)	<0.001
Diabetes, n (%)	6425 (30.0)	4159 (15.6)	<0.001
Hypertension, n (%)	6625 (34.2)	8349 (33.3)	0.065
Current smoking, n (%)	10254 (46.2)	6916 (24.9)	<0.001
LDL cholesterol* (mg/dL), mean (SD)	135.6 (53.9)	98.7 (43.2)	<0.001
HDL cholesterol (mg/dL), mean (SD)	36.4 (11.7)	35.2 (14.3)	<0.001
Triglycerides† (mg/dL), median [Q1,Q3]	163.0 [111.0,241.0]	107.0 [31.1,187.0]	<0.001

The Q1 and Q3 are the first and third quartiles of the distribution. LDL, low-density lipoprotein. HDL, high-density lipoprotein. SD, standard deviation. *Estimated untreated LDL cholesterol, divided value by 0.70 if taking lipid-lowering medicine assuming statin. † Estimated untreated Triglycerides, divided value by 0.85 if taking lipid-lowering medicine assuming statin.

Table II. Study design and definitions of coronary artery disease among case-control cohorts of the Myocardial Infarction Genetics ExSeq study

Study	N Cases	N Controls	Ancestry	Capture Kit	Coronary Artery Disease Definition	Control Definition	Ref
ARIC	725	8220	European African	Nimblegen	Incident probable or definite MI, silent MI, definite CAD death, or coronary revascularization	Free of CAD during follow-up	57
ATVB	1803	1725	European	Agilent	MI in men or women ≤ 45 years of age	No history of cardiovascular or thromboembolic disease	58
BRAVE	746	739	South Asian	Nextera	MI in men and women ≤60 years	Controls without CAD; men and women ≤65 years	43,59
EOMI	994	1718	European African	Agilent	MI in men ≤ 50 years of age or women ≤ 60 years of age	Free of MI, coronary revascularization; men ≥ 50 years of age or women ≥ 60 years of age	60
Leicester	1175	1097	European African South Asian	ICE	MI in men or women age ≤60 years	Controls ≥ 64 years without reported CAD history	61
North German MI	867	873	European	ICE	MI in men and women ≤ 60 years	Controls without CAD; men and women ≤ 65 years	62
OHS	970	977	European	Agilent	Angiographically confirmed coronary artery disease (>1 coronary artery with >50% stenosis) without a history of diabetes at age ≤ 50 for men or ≤ 60 for women	Asymptomatic, men > 65, women > 70	63
PROCARDIS	975	962	European	Agilent	Symptomatic CAD before age 66. CAD was defined as clinically documented evidence of myocardial infarction, coronary artery bypass grafting, acute coronary syndrome, coronary angioplasty, or stable angina	No personal or sibling history of CAD before 66 years of age	64
PROMIS	15074	13253	South Asian	Agilent/ICE/Nextera	Myocardial infarction	No history of cardiovascular disease	65,66
REGICOR	368	392	European	ICE	MIinmen≤50yearsofageorwomen≤60yearsof age	Controls from a population-based study; free of MI, coronary revascularization; ≥ 55 and <80 years of age	67
South German MI	400	398	European	ICE	MI in men ≤ 40 years of age or women ≤ 55 years of age	Controls without CAD, men ≥ 65 years of age and women ≥ 75 years of age	68

ARIC: The Atherosclerosis Risk in Communities (ARIC) Study.[57] ATVB: the Atherosclerosis Thrombosis and Vascular Biology (ATVB) study.[58] BRAVE : The Bangladesh Risk of Acute Vascular Events. EOMI: the Exome Sequencing Project Early-Onset Myocardial Infarction (ESP-EOMI) study.[60] Leicester: the Leicester Myocardial Infarction study.[61] North German MI: the North German Myocardial Infarction study.[62] OHS: the Ottawa Heart Study.[63] PROCARDIS: the Precocious Coronary Artery Disease Study.[64] PROMIS: the Pakistan Risk of Myocardial Infarction Study.[65,66] REGICOR: the Registre Gironi del COR (Gerona Heart Registry or REGICOR) study.[67] South German MI: the South German Myocardial Infarction study.[68]

Nimblegen, Nimblegen exome capture array (HGSC VCRome 2.1 design). Agilent, Agilent SureSelect Human All Exon. Nextera, Illumina Nextera Exome Kit. ICE, Illumina ICE exome capture kit.

Table III. Clinical characteristics for Myocardial Infarction Genetics WGSeq study.

VARIABLE	CASE[‡]	CONTROL	P-Value
N	2369	4218	
Censor age, mean (SD)	48.3 (6.4)	61.3 (9.8)	<0.001
Sex, Male, n (%)	925 (39.0)	2019 (47.9)	<0.001
Ancestry, n (%)			
African	336 (14.2)	962 (22.8)	<0.001
East Asian	328 (13.8)	961 (22.8)	
European	1537 (64.9)	1544 (36.6)	
Other	168 (7.1)	751 (17.8)	
Lipid-lowering therapy, n (%)	668 (28.7)	584 (13.9)	<0.001
Diabetes, n (%)	876 (37.1)	665 (15.8)	<0.001
Hypertension, n (%)	1415 (59.9)	1600 (37.9)	<0.001
Current smoking, n (%)	1146 (48.9)	535 (12.7)	<0.001
LDL cholesterol* (mg/dL), mean (SD)	122.1 (47.5)	122.5 (34.5)	0.778
HDL cholesterol (mg/dL), mean (SD)	40.5 (13.4)	51.0 (15.0)	<0.001
Triglycerides† (mg/dL), median [Q1,Q3]	142.7 [98.0,220.2]	115.0 [79.0,166.0]	<0.001

The Q1 and Q3 are the first and third quartiles of the distribution. LDL, low-density lipoprotein. HDL, high-density lipoprotein. SD, standard deviation*Estimated untreated LDL cholesterol, divided value by 0.70 if taking lipid-lowering medicine assuming statin. † Estimated untreated Triglycerides, divided value by 0.85 if taking lipid-lowering medicine assuming statin. ‡ 2081 cases from the VIRGO study were enrolled by a 2:1 female-to-male design[5].

Table IV. Clinical characteristics for UK Biobank 13K study.

VARIABLE	CASE	CONTROL	P-Value
N	6446	5932	
CAD onset age, mean (SD)	50.5 (7.9)	NA	
Enrollment age, mean (SD)	58.8 (7.2)	58.7 (7.1)	0.649
Sex, Male, n (%)	4205 (65.2)	3894 (65.6)	0.646
Ancestry, n (%)			
African	72 (1.1)	56 (0.9)	<0.001
East Asian	10 (0.2)	13 (0.2)	
European	5980 (92.8)	5718 (96.4)	
Other	137 (2.1)	77 (1.3)	
South Asian	247 (3.8)	68 (1.1)	
Lipid-lowering therapy, n (%)	4684 (72.7)	1068 (18.0)	<0.001
Diabetes, n (%)	1339 (20.8)	226 (3.8)	<0.001
Hypertension, n (%)	4568 (70.9)	1842 (31.1)	<0.001
Current smoking, n (%)	1246 (19.5)	520 (8.8)	<0.001
Family history of heart disease, n (%)	4059 (63.0)	2464 (41.5)	<0.001
BMI, mean (SD)	29.6 (5.4)	27.3 (4.3)	<0.001
LDL cholesterol* (mg/dL), mean (SD)	149.0 (37.9)	147.1 (31.6)	0.003
HDL cholesterol* (mg/dL), mean (SD)	47.3 (13.2)	54.4 (14.3)	<0.001
Triglycerides* (mg/dL), median [Q1,Q3]	174.3 [120.8,251.3]	141.8 [99.9,205.5]	<0.001

The Q1 and Q3 are the first and third quartiles of the distribution. SD, standard deviation. LDL, low-density lipoprotein. HDL, high-density lipoprotein. *Estimated untreated lipid levels[54].

Table V. Clinical characteristics for UK Biobank 200K study.

VARIABLE	CASE	CONTROL	P-Value
N	8169	176611	
CAD onset age, mean (SD)	62.3 (7.6)	NA	
Enrollment age, mean (SD)	62.8 (5.7)	56.7 (8.0)	<0.001
Sex, Male, n (%)	6492 (79.5)	77120 (43.7)	<0.001
Ancestry, n (%)			
African	50 (0.6)	3011 (1.7)	<0.001
East Asian	12 (0.1)	610 (0.3)	
European	7671 (93.9)	165389 (93.6)	
Other	147 (1.8)	3848 (2.2)	
South Asian	289 (3.5)	3753 (2.1)	
Lipid-lowering therapy, n (%)	4941 (60.5)	30270 (17.1)	<0.001
Diabetes, n (%)	1515 (18.5)	7428 (4.2)	<0.001
Hypertension, n (%)	5824 (71.3)	53634 (30.4)	<0.001
Current smoking, n (%)	1026 (12.7)	16659 (9.5)	<0.001
Family history of heart disease, n (%)	4757 (58.3)	73851 (41.9)	<0.001
BMI, mean (SD)	28.8 (4.6)	27.3 (4.7)	<0.001
LDL cholesterol* (mg/dL), mean (SD)	149.3 (36.6)	145.4 (33.0)	<0.001
HDL cholesterol* (mg/dL), mean (SD)	47.8 (12.4)	56.7 (14.8)	<0.001
Triglycerides* (mg/dL), median [Q1,Q3]	169.6 [118.8,244.2]	132.9 [93.4,193.4]	<0.001

The Q1 and Q3 are the first and third quartiles of the distribution. SD, standard deviation. LDL, low-density lipoprotein. HDL, high-density lipoprotein. *Estimated untreated lipid levels[54].

Table VI. Biomarker association for NOS3 in UK Biobank 200K.

	NONCARRIERS			CARRIERS					
TRAIT (unit)	MEAN	SD	N	MEAN	SD	N	Beta	SE	P-value
LDL cholesterol* (mg/dL)	145.5	32.6	174,692	146.3	35.5	831	+ 0.8	1.1	0.47
HDL cholesterol* (mg/dL)	56.3	13.5	160,937	56.0	13.2	766	- 0.2	0.5	0.64
Total cholesterol* (mg/dL)	228.2	41.2	174,998	227.8	43.1	833	- 0.4	1.4	0.76
Triglycerides* (mg/dL)	159.2	92.7	174,852	159.0	83.0	833	- 0.2	3.2	0.94
Systolic blood pressure† (mmHg)	143.7	20.7	173,692	147.0	22.2	850	+ 3.3	0.7	0.000005
Diastolic blood pressure† (mmHg)	84.2	11.6	173,697	85.5	12.2	850	+ 1.3	0.4	0.0007

The mean and SD (standard deviation) were estimated using the covariate standardized values. The Beta, SE, and *P-value* were estimated for the effect of carrier status on each biomarker in a regression model adjusted the covariates. The covariates used for the adjustment are enrollment age, sex, and top 10 principal components of ancestry. SD, standard deviation. SE, standard error. *Estimated untreated lipid levels[54]. † Estimated untreated blood pressure levels, adding 15 mm Hg to systolic and 10 mm Hg to diastolic values, respectively, to participants who reported use of blood-pressure lowering medications[55,56].

Table VII. Rare variant annotations for *NOS3*, *GUCY1A3* and *GUCY1B3*.

TableVII. Rare variant annotations for NOS3, GUCY1A3 and GUCY1B3.xlsx

LOF, LOFTEE algorithm identified ‘high-confidence’ stop-gain, splice-site disrupting, and frameshift variants.

OF5, ultra-rare missense variants predicted to be damaging by each of five computational prediction algorithms.

SAI, SpliceAI algorithm identified cryptic splice sites.

Table VIII. Top 15 genes identified by the 'pLoF+missense' strategy.

Gene	Chromosome	N Cases	N Controls	Beta	SE(Beta)	P value
LDLR	19	375	737	1.48	0.12	1.74E-32
NOS3	7	242	983	0.88	0.15	5.50E-09
CTD-3214H19.16	19	8	16	2.4	0.52	3.31E-06
LPIN2	18	141	646	0.78	0.19	4.99E-05
MED28	4	61	194	1.09	0.28	1.37E-04
TEC	4	100	384	1.03	0.27	1.38E-04
CCDC27	1	62	447	-0.72	0.19	1.38E-04
GOT1	10	64	251	1.05	0.29	2.20E-04
C16orf74	16	17	78	2.41	0.66	2.85E-04
INIP	9	13	37	1.42	0.39	2.92E-04
PANX1	11	90	375	1	0.28	3.42E-04
C18orf21	18	48	134	0.88	0.25	3.61E-04
HTRA1	10	72	220	0.89	0.25	4.01E-04
ZNF687	1	37	99	0.95	0.27	4.21E-04
MADCAM1	19	11	64	1.41	0.4	4.42E-04

SE: standard error. N cases: the number of carriers identified from 41,081 cases. N Controls: the number of carriers identified from 217,115 controls.

Table IX. Test the NOS3 association by using alternate weights for the SpliceAI variants.

Weight	OR	95% CI	P value	Heterogeneity(Q) P-Value *
0.75	2.42	1.80 - 3.26	5.5 x 10-9	
0.5	2.55	1.87 - 3.47	3.1 x 10-09	0.81
1.0	2.26	1.70 - 2.99	1.5 x 10-08	0.74

* Compared with the results of using weight of 0.75.

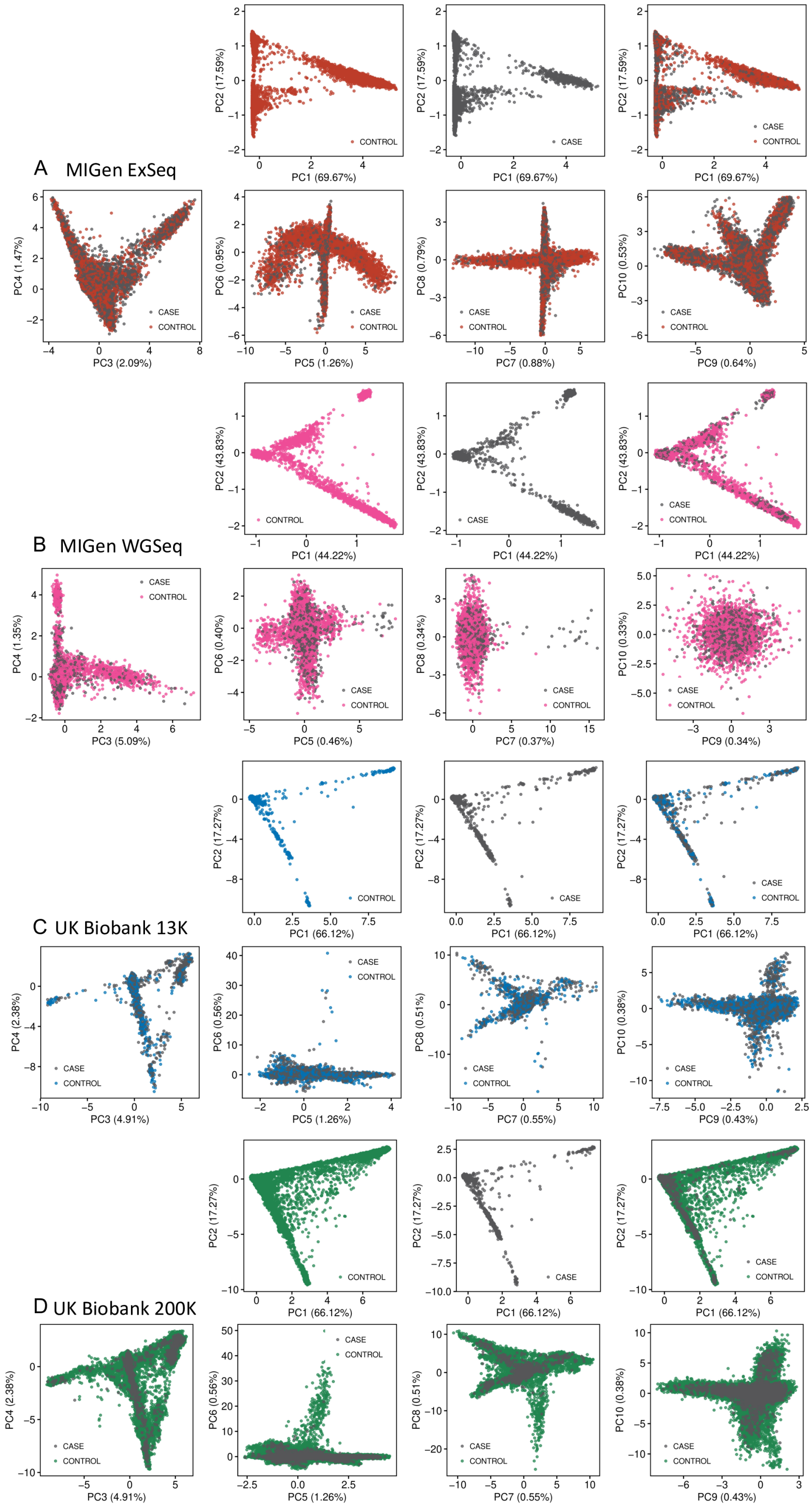
Table X. The covariates adjusted in each study.

STUDY	COVARIATES
MIGen ExSeq	Top 10 PCs + Sex + individual study name
UK Biobank 13K	Top 10 PCs + Sex
UK Biobank 200K	Top 10 PCs + Sex + Enrollment age
MIGen WGSeq	Top 10 PCs

PCs: principal components.

Supplementary Figures

Figure I. Principal components of ancestry for study participants across four cohorts.

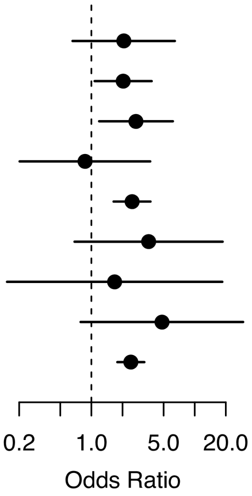


Genetic ancestry plot for the study samples by principal component analysis stratified by case and control status. **A**) Principal components of all of the Myocardial Infarction Genetics ExSeq study participants. **B**) Principal components of the Myocardial Infarction Genetics WGSeq study participants. **C**) Principal components of the UK Biobank 13K Study participants. **D**) Principal components of the UK Biobank 200K Study participants. The percentage in each parenthesis was the proportion of variance explained by each principal component of the total variance captured by the top 20 principal components.

Figure II. Association of NOS3 gene rare variants with coronary artery disease risk by genetic ancestry.

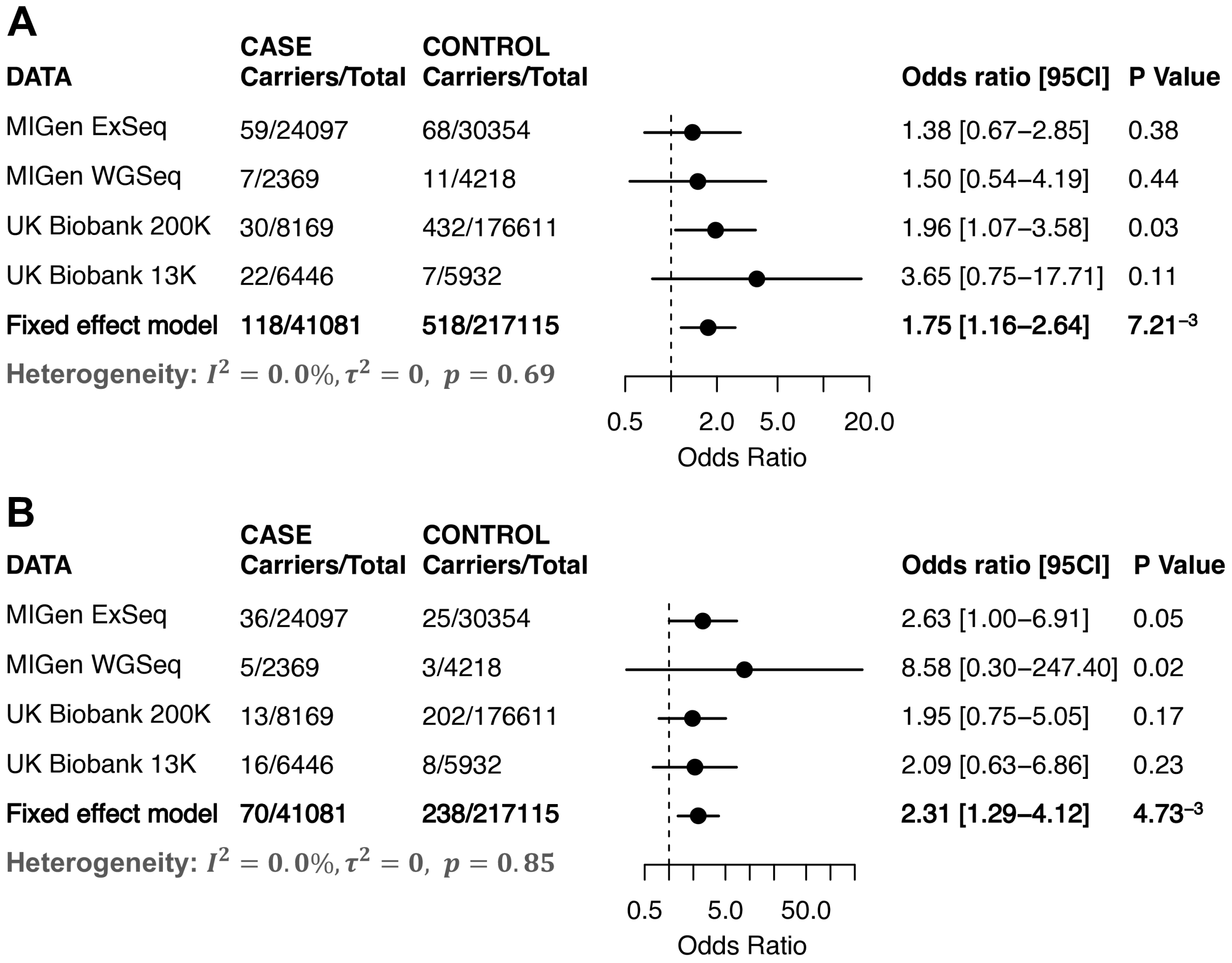
DATA	CASE Carriers/Total	CONTROL Carriers/Total	Odds ratio [95CI]	P Value
MIGen ExSeq_EUR	33/7608	51/13448	2.05 [0.66–6.42]	0.216
MIGen ExSeq_nonEUR	88/16489	72/16906	2.03 [1.08–3.83]	0.029
UK Biobank 13K_EUR	35/5980	18/5718	2.70 [1.19–6.14]	0.018
UK Biobank 13K_nonEUR	6/466	3/214	0.87 [0.20–3.71]	0.846
UK Biobank 200K_EUR	61/7671	724/165389	2.47 [1.64–3.73]	0.151^{-06}
UK Biobank 200K_nonEUR	4/498	99/11222	3.58 [0.69–18.61]	0.130
MIGen WGSeq_EUR	10/1537	7/1544	1.68 [0.15–18.47]	0.671
MIGen WGSeq_nonEUR	5/832	9/2674	4.81 [0.79–29.37]	0.089
Fixed effect model	**242/41081**	**983/217115**	**2.41 [1.78–3.25]**	**0.953^{-10}**

Heterogeneity: $I^2 = 0.0\%, \tau^2 = 0,\ p = 0.89$



The samples in each data set were partitioned into two groups according to European (EUR) or non-European (nonEUR) ancestry, the association was then tested in each group used the same variant as for the main analysis. The inverse variance weighted fixed-effect meta-analysis was applied to summarize the results.

Figure III. Forest plot for gene GUCY1A3 and GUCY1B3.



A

DATA	CASE Carriers/Total	CONTROL Carriers/Total	Odds ratio [95CI]	P Value
MIGen ExSeq	59/24097	68/30354	1.38 [0.67–2.85]	0.38
MIGen WGSeq	7/2369	11/4218	1.50 [0.54–4.19]	0.44
UK Biobank 200K	30/8169	432/176611	1.96 [1.07–3.58]	0.03
UK Biobank 13K	22/6446	7/5932	3.65 [0.75–17.71]	0.11
Fixed effect model	**118/41081**	**518/217115**	**1.75 [1.16–2.64]**	**7.21^{-3}**

Heterogeneity: $I^2 = 0.0\%, \tau^2 = 0,\ p = 0.69$

B

DATA	CASE Carriers/Total	CONTROL Carriers/Total	Odds ratio [95CI]	P Value
MIGen ExSeq	36/24097	25/30354	2.63 [1.00–6.91]	0.05
MIGen WGSeq	5/2369	3/4218	8.58 [0.30–247.40]	0.02
UK Biobank 200K	13/8169	202/176611	1.95 [0.75–5.05]	0.17
UK Biobank 13K	16/6446	8/5932	2.09 [0.63–6.86]	0.23
Fixed effect model	**70/41081**	**238/217115**	**2.31 [1.29–4.12]**	**4.73^{-3}**

Heterogeneity: $I^2 = 0.0\%, \tau^2 = 0,\ p = 0.85$

Rare variant burden test aggregated variants predicted to lead to loss-of-function, disrupt mRNA splicing, or annotated as pathogenic or likely pathogenic within the ClinVar database, or ultra-rare missense variants predicted to be damaging by each of five computational prediction algorithms. The meta-analysis was performed by a fixed-effects meta-analysis model based on the effect size estimated from a Firth logistic regression in each of the four studies. Panel **a**) is a forest plot including carrier counts across cases and controls within four studies for the gene Guanylate Cyclase Soluble Subunit Alpha-3 (*GUCY1A3*). Panel **b**) is the forest plot results for the gene Guanylate Cyclase Soluble Subunit Beta-3 (*GUCY1B3*). The bar in both plots presents 95% confidence interval.