Journal of the American Heart Association

ORIGINAL RESEARCH

Evolutionary Action–Machine Learning Model Identifies Candidate Genes Associated With Early-Onset Coronary Artery Disease

Dillon Shapiro, BS; Kwanghyuk Lee, PhD; Jennifer Asmussen, PhD; Thomas Bourquard, PhD; Olivier Lichtarge, MD, PhD

BACKGROUND: Coronary artery disease is a primary cause of death around the world, with both genetic and environmental risk factors. Although genome-wide association studies have linked >100 unique loci to its genetic basis, these only explain a fraction of disease heritability.

METHODS AND RESULTS: To find additional gene drivers of coronary artery disease, we applied machine learning to quantitative evolutionary information on the impact of coding variants in whole exomes from the Myocardial Infarction Genetics Consortium. Using ensemble-based supervised learning, the Evolutionary Action–Machine Learning framework ranked each gene's ability to classify case and control samples and identified 79 significant associations. These were connected to known risk loci; enriched in cardiovascular processes like lipid metabolism, blood clotting, and inflammation; and enriched for cardiovascular phenotypes in knockout mouse models. Among them, *INPP5F* and *MST1R* are examples of potentially novel coronary artery disease risk genes that modulate immune signaling in response to cardiac stress.

CONCLUSIONS: We concluded that machine learning on the functional impact of coding variants, based on a massive amount of evolutionary information, has the power to suggest novel coronary artery disease risk genes for mechanistic and therapeutic discoveries in cardiovascular biology, and should also apply in other complex polygenic diseases.

Key Words: coronary artery disease ■ evolutionary action ■ gene-based associations ■ machine learning ■ myocardial infarction

Coronary artery disease (CAD) remains the global leading cause of death. Greater than 20 million adults are diagnosed with CAD, and approximately 650 000 people in the United States die annually from some form of heart disease.[1,2] CAD stems from a combination of genetic and environmental factors; therefore, understanding which individuals are most at risk for the development of disease could facilitate earlier lifestyle and pharmacological interventions. Past studies have estimated genetic heritability for CAD to be between 40% and 60%[3,4] and identified >100 loci associated with CAD and its related phenotypes, including variants primarily related to cholesterol metabolism, vascular remodeling, inflammation, and angiogenesis.[5–7] Recent population scale analyses with hundreds of thousands of samples such as the coronary artery disease genome wide replication and meta-analysis plus the coronary artery disease genetics consortium,[8,9] UK Biobank,[10,11] and the trans-omics for precision medicine program[12] have replicated the association of several previously reported common variants. However, these associations often have uncertain implications due to their frequent localization in noncoding regions of the genome. As a complement to common variant associations, whole exome sequencing studies[13–15] have identified

Correspondence to: Olivier Lichtarge, MD, PhD, Baylor College of Medicine, One Baylor Plaza, T921, Houston, TX 77030. Email: lichtarg@bcm.edu

This article was sent to Julie K. Freed, MD, PhD, Associate Editor, for review by expert referees, editorial decision, and final disposition.

Supplemental Material is available at https://www.ahajournals.org/doi/suppl/10.1161/JAHA.122.029103

For Sources of Funding and Disclosures, see page 13.

JAHA is available at: www.ahajournals.org/journal/jaha

CLINICAL PERSPECTIVE

What Is New?

- A novel exome-wide association method using evolutionary information and machine learning reveals multiple genes with functional evidence linked to early-onset cardiovascular disease.
- Criteria for successful experiments support contributions from known and suggested coronary artery disease–associated risk genes in important biological pathways.

What Are the Clinical Implications?

- The protein-coding variation characterized in this study further expands our knowledge of genetic contributions to coronary artery disease risk and prioritizes new avenues for mechanistic interrogations, polygenic risk assessment, and therapeutics.

Nonstandard Abbreviations and Acronyms

EA	Evolutionary Action
EAML	Evolutionary Action–Machine Learning
EOMI	early-onset myocardial infarction
MCC	Matthew correlation coefficient
MIGen	Myocardial Infarction Genetics
STARNET	Stockholm-Tartu Atherosclerosis Reverse Network Engineering Task

rare mutations with the strongest links to heart disease risk and age of onset, including lipid metabolism genes *LDLR*, *LDLRAP1*, *APOB*, and *PCSK9*, which are causal for familial hypercholesterolemia and are associated with dramatic increases in low-density lipoprotein cholesterol.[16–20] However, rare variant studies are typically underpowered due to sample sizes that are magnitudes smaller than those used in traditional array-based approaches.[21] Consequently, fewer variants and genes from rare variant-based approaches meet the statistical significance threshold to be of interest for follow-up analyses, leaving a gap in the identification of novel CAD risk factors.

Although current genome-wide association studies (GWAS) findings explain approximately 40% to 50% of estimated heritability in CAD, with common and rare variants accounting for ≈20% and ≈2% to 4%, respectively,[22,23] one possible source of missing heritability is the potential for variants to have nonadditive effects on disease risk[24] based on their functional impact. Standard modeling techniques rely on strict, linear assumptions about genetic inheritance,[25–28] and approaches that consider nonlinear and nonadditive interactions have shown improvements in some cases.[29–31] Although recent studies have begun to explore the impact of variation beyond the level of independent single nucleotide polymorphisms (ie, rare variant burden, epistatic interactions between genes/variants, network interactions),[25,26,32] association methods could be augmented by the fact that variants have different levels of impact on protein structure and function across evolution. Together, this argues that new approaches are needed to identify more direct, functional connections between genetic factors and CAD.

To address these issues, we developed a novel Evolutionary Action–Machine Learning (EAML) framework that scores the relative contribution of a gene's mutations in distinguishing individuals affected by a complex trait from healthy controls.[33] By incorporating the Evolutionary Action (EA) functional impact score,[34] derived from a systematic analysis of evolutionary information on protein sequence variations and divergences,[35] into association testing and focusing on protein-coding variants, we include a larger portion of variance in our association modeling that is directly related to biological importance. The use of EA has been previously demonstrated in blinded community challenges to detect deleterious coding variants[36] and in identifying genes associated with Alzheimer disease,[33,37,38] cancer,[32,39–41] autism spectrum disorder,[42] and antibiotic resistance.[43] Additionally, the use of machine learning allows us to address nonlinear patterns of variation within potential risk genes. Here, we have used EAML to search for novel risk genes in early-onset myocardial infarction (EOMI), an outcome of CAD where genetic inheritance is a major risk component. First, we flagged potential risk genes using EAML on 7426 samples with EOMI and healthy samples from the Myocardial Infarction Genetics (MIGen Consortium).[13] We then assessed these EAML candidates against known CAD-related risk genes and traits and further characterized them through clustering with known risk genes in a protein–protein interaction network. Finally, we prioritized EAML candidates by aggregating evidence related to GWAS, relative risk, mouse knockout data, expression quantitative trait loci (eQTLs), and PubMed co-occurrences. EAML recovered the most important known biological associations through direct overlap and pathway enrichments, but also prioritized novel candidates through multiple computational criteria. These results suggest an increasingly important role for genes that regulate lipid metabolism, inflammation, blood clotting, and the cell cycle and open new directions for mechanistic and therapeutic research in CAD.

METHODS

Data Disclosure

This study protocol (H-37394) was approved by the Institutional Review Board for Human Subject Research for Baylor College of Medicine and Affiliated Hospitals.

Our analyses were based on the following 3 data sets: the ATVB (Atherosclerosis, Thrombosis, and Vascular Biology) study, the OHS (Ottawa Heart Study), and the PROCARDIS (Precocious Coronary Artery Disease) study. These studies were approved by the institutional review boards of all participating institutions, and information about informed consent from the study participants can be found in the study homepages (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000814.v1.p1, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000883.v1.p1, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000806.v1.p1).

Because of the sensitive nature of the data collected for this study, requests to access the data sets from qualified researchers may be sent to the database of Genotypes and Phenotypes (dbgap-help@ncbi.nlm.nih.gov).

Software Availability

EA scores of missense variants are publicly available via web server (http://eaction.lichtargelab.org/). EAML code is publicly available on GitHub (https://github.com/LichtargeLab/EAML), including easy installation and a toy data set for testing.

Data Collection

Whole-exome sequencing data were obtained from 3 previously established studies within the MIGen Exome Sequencing Consortium: the ATVB study (dbGaP accession: phs000814),[44] the OHS (dbGaP accession: phs000806),[45] and the PROCARDIS study (dbGaP accession: phs000883).[46] In ATVB, both cases and controls were selected from across 125 coronary care units in Italy between 1994 and 2007. Cases were defined as patients hospitalized for a first myocardial infarction (MI) <45 years of age, and controls were age- and sex-matched individuals without a history of thromboembolic disease, a subclass of cardiovascular disease that shares many of the underlying risk factors with CAD. In OHS, cases were selected through the Ottawa Heart Institute and defined as patients diagnosed with CAD (defined as MI, coronary artery bypass, or angiographic stenosis >50%) <55 years of age in men and <65 years of age in women, whereas controls were men >65 years of age and women >70 years of age without a history of cardiovascular disease and selected through newspaper and television advertising. In PROCARDIS, cases were selected from hospitals in the United Kingdom and defined as patients with MI, unstable or stable angina, or coronary revascularization <66 years of age. Controls were age- and sex-matched individuals without personal or sibling history of cardiovascular disease, recruited from the same centers through self-administered questionnaires. The cardiovascular disease history of the control samples is defined according to the case definition in each respective study. Available clinical characteristics of each cohort were limited to those provided by the studies in dbGaP, including case–control numbers and reported sex distributions. The age cutoffs and sample selection criteria were predetermined by the individual studies, and detailed descriptions of sample selection criteria are available in each cohort's original publication.

Variant Quality Control

To remove potentially low-quality variant sites, we filtered variant sites based on the average genotype quality, average depth of coverage, missingness, and Hardy-Weinberg Equilibrium test. Variants with an average genotype quality <20, average depth of coverage <8, missingness >2%, or Hardy-Weinberg Equilibrium *P* value of controls $<5\times10^{-5}$ were removed from the data set. All variant filtering steps were performed using the BCFtools software.[47]

Sample Quality Control

We performed multiple steps to identify outlier samples before analysis. First, we inferred ancestry by using principal component analysis to map samples with 1000 genomes and excluded samples that were inferred as non-European (Figure S1). Second, we removed samples with an excess of coding variants (>17 000), an excess of missing variants (>1300), or an excess of singletons (>20), because these can suggest low-quality sequencing (Figure S2). We carefully considered these thresholds through the visualization of sample statistics. Next, we inferred sex with the ratio of heterozygous to homozygous variants on the X chromosome and removed samples with a mismatch between inferred and self-reported sex (Figure S3). Finally, we estimated kinship coefficients between samples and removed samples with third-degree relatives (kinship coefficient >0.1; Figure S4). Principal component analysis, sex, and relatedness filtering were performed using the *peddy* Python package.[48]

Variant Annotation

Variants (single nucleotide variants and indels) were annotated using the hg19 RefSeq reference and the Annotate Variation annotation tool.[49] Nonsynonymous variants were annotated with the EA equation,[34] receiving a variant impact score between 0 and 100. We assigned loss-of-function variants such as frameshift indels and stop-gain variants a maximal EA score of 100.

Statistical Analysis

Rare Variant Association Analyses

As a control experiment and to replicate previous studies, we performed a rare variant burden test using the optimal sequence association test method[25] in the efficient and parallelizable association container toolbox package (https://genome.sph.umich.edu/wiki/EPACTS). We selected variants with a minor allele frequency <1% and collapsed variants across a gene using EA impact score thresholds to compare optimal sequence association test performance more directly to EAML. We used 3 different variant groupings: (1) all nonsynonymous variants, (2) partially deleterious variants based on having an EA score >30, and (3) deleterious variants based on an EA score>70.

EAML Pipeline

Quantifying Functional Genetic Burden

For each gene, we first calculated an aggregate score of all variant level effects into 1 metric for 6 underlying genotype–phenotype association hypotheses. For this, we developed a function dubbed the EA probability (pEA). This is defined as:

$$pEA = 1 - \prod_{j=1}^{k}\left(1 - \frac{EA_j}{100}\right)^{zyg} \quad \forall EA > C \qquad (1)$$

where k is the total number of variants in a gene, j is the index over those variants, EA_j is the EA score from a given variant, and zyg denotes the zygosity of variant j (0 denotes wild-type, 1 denotes heterozygous, and 2 denotes homozygous). C denotes 3 different thresholds of EA, specifically 0, 30, and 70. The 6 underlying hypotheses are delineated by terms C and zyg. First, the thresholds defined by C correspond to how the degree of predicted variant impact associates with disease status: (1) any missense variant (EA >0), (2) moderate-to-high impact variants (EA >30), or (3) deleterious variants (EA >70). Second, the *zyg* term avoids a priori assumptions about a gene's inheritance pattern, allowing for an association in either an autosomal dominant (zyg >0) or recessive (zyg >1) manner. Mutations are separated into 6 pEA learning features based on the assumptions made by C and zyg. Finally, these features are aggregated into a nx*6* design matrix for each gene, where n is the number of samples.

Model Architecture

The learning architecture consisted of 9 different classifiers, representing standard models used in modern machine learning problems, combined in an averaging ensemble. These classifiers include Association Rules (PART,[50] JRip [51]), Function Optimizations (Multilayer Perceptron,[52] Naïve Bayes,[53] Logistic Regression,[54] and K Nearest Neighbors[55]), Decision Trees (Random Forest[56] and J48[57]), and meta-classifiers (Adaboost[58]). All classifiers were implemented in Weka with default hyperparameters (https://www.cs.waikato.ac.nz/ml/weka/).

Association Testing

To evaluate the association of each gene, each member of the 9-classifier ensemble was trained on 90% of the input data set to classify disease cases from healthy controls. Performance was evaluated on the leftover 10%, generating a classification score for a given gene. This process was repeated in a 10-fold cross validation, with the scores averaged across folds, to minimize data set overfitting.

The classification performance of an individual gene was used as a surrogate for the magnitude of association with the given phenotype. This was estimated using the Matthew correlation coefficient (MCC), defined as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \qquad (2)$$

The MCC scores from all 9 classifiers were averaged and used to produce a final ranking of all genes. We then computed Z scores and the corresponding 1-tailed Z test P values for each average MCC relative to the full distribution of gene MCC scores. Finally, we implemented a false discovery rate correction using the Benjamini-Hochberg method, and genes with a false discovery rate–corrected P value <0.1 were selected as candidate risk genes. More details of the EAML approach can be found in Data S2 (Figure S5).

Mouse Phenotype Analysis

To assess biological causality between EAML genes and cardiovascular phenotypes, we queried the Mouse Genome Informatics database of assayed mouse knockout models.[59] To test whether EAML candidates are linked to cardiovascular phenotypes more often than random genes, we counted the number of candidates with a mouse model reporting a cardiovascular system phenotype. Then, we randomly sampled gene sets of the same size and repeated the counting procedure to generate a random background distribution. We then calculated the Z score and P value of the candidate gene list's enrichment in comparison with the random background distribution.

STRING Network Analyses

All protein–protein interaction networks were constructed using the STRING protein-protein interaction database version 11.0, downloaded from https://version-11-0b.string-db.org/. Edges with a combined score (from all evidence types)>400 were included in network analyses.

Graph information diffusion[60–62] was used to assess how closely related EAML genes are to target gene sets within a protein–protein interaction network. Graph information diffusion propagates signal from an input gene set across the network and reported the signal received by every gene. The closeness of the input and target gene sets was then reported as a receiver operating characteristic and the area under the curve (AUC). Because there is a large class imbalance with many more genes labeled as unrelated to MI, AUC can be an inflated measure of performance. To control for this, we randomly sampled gene sets with the same size and degree distribution as the target gene set and repeated the graph information diffusion analyses. AUCs were measured for these random gene sets, then used as a background distribution for which the *Z* score of the true target gene set was measured. A *Z* score >1.96 (*P*<0.05) suggested that the input genes and target genes were functionally related.

Odds Ratio Calculation

To identify the directionality of each candidate gene's/variant's contribution to relative risk, which is not revealed by EAML, we calculated a crude allelic odds ratio for each candidate gene and each protein-coding variant within each candidate gene. Statistical significance was calculated using the Fisher exact test. We corrected for multiple testing using the Benjamini-Hochberg false discovery rate correction and assigned statistical significance with an adjusted *P* value threshold of 0.01.

Gene-Level eQTL Mapping

Data on healthy tissue eQTLs were acquired from the Genotype-Tissue Expression (GTEx) Project database.[63] We considered any eQTL reported in any tissue with a *P* value $<1\times10^{-6}$. Cardiovascular eQTLs were identified using the STARNET (Stockholm-Tartu Atherosclerosis Reverse Network Engineering Task) study.[64] The data from this study contain both normalized RNA expression and genotyped DNA across 7 cardiovascular tissue types. We used a set of eGenes reported by the initial study, which are genes containing at least 1 *cis*-eQTL.

RESULTS

EAML Identified 79 EOMI-Associated Disease Candidate Genes Using Evolution-Based Machine Learning

To discover novel risk genes associated with EOMI using evolutionary information, we aggregated whole exome sequencing data from 3 cohorts within the MIGen (Table S1). These data included 3736 individuals diagnosed with EOMI who also underwent coronary angiography. The controls included 3690 healthy subjects with no history of thromboembolic or cardiovascular disease. After annotating the cohort variants with gene, transcript, and EA scores, we performed quality control to exclude potentially false-positive variants as well as individuals of non-European descent, mismatched sex, and sequencing outliers (Figure 1A; see Methods). We then analyzed the cohorts with EAML, an ensemble-based pipeline that evaluates each gene's ability to classify cases from controls using EA scores and supervised machine learning.

For each gene and individual within the cohort, EAML first calculates a probability of functional impact by aggregating EA scores for 6 different groups of variants, each associated with their own unique underlying hypothesis (Figure 1B). We defined the groups by the EA magnitude and inheritance pattern, allowing EAML to selectively evaluate the importance of each gene based on different patterns of functional impact and variation. Each gene-based feature matrix was then used as the input for an ensemble of supervised classifiers, and the average MCC[65] score across classifiers was calculated for each gene. Averaging was used to determine the consensus among model types while also reducing false positives and focusing on specificity over sensitivity. Finally, each gene was ranked and prioritized based on its average MCC score. For our analysis of the EOMI cohort, we applied EAML in an unbiased fashion, using all variants scored with EA regardless of allele frequency. Using EAML, we identified 79 genes (Table S2) passing the false discovery rate–corrected *P* value threshold of 0.1 and having a positive MCC score (Figure 1C) after 10-fold cross-validation on 16912 genes with non-0 discriminatory power (Figure S6A and S6B). For comparison, we performed a rare variant association analysis, optimal sequence association test, on the same data and identified a single gene meeting statistical significance ($P<5\times10^{-6}$), the widely known CAD risk gene *LDLR* (Figure S6C). These data show how EAML prioritizes potential EOMI risk genes using evolutionary information and ensemble machine learning, recovering more genes for computational and experimental validation than current state-of-the-art association methods. The remainder of the study was focused on these 79 candidate genes.

EAML Candidates Are Enriched in Cardiovascular Disease Gene Sets and Related Phenotypes

To assess the ability of EAML to recover genes associated with cardiovascular phenotypes and their biomarkers, we first tested the 79 EOMI candidate genes for phenotype enrichment within the GWAS Catalog database using the functional mapping and annotation portal.[66] We found significant enrichments for CAD ($P=5.20\times10^{-11}$) and MI ($P=3.48\times10^{-6}$) as well as for multiple lipid biomarkers and related phenotypes (Figure 2B), which included triglyceride:high-density lipoprotein (HDL) ratio ($P=9.63\times10^{-5}$), hypertriglyceridemia

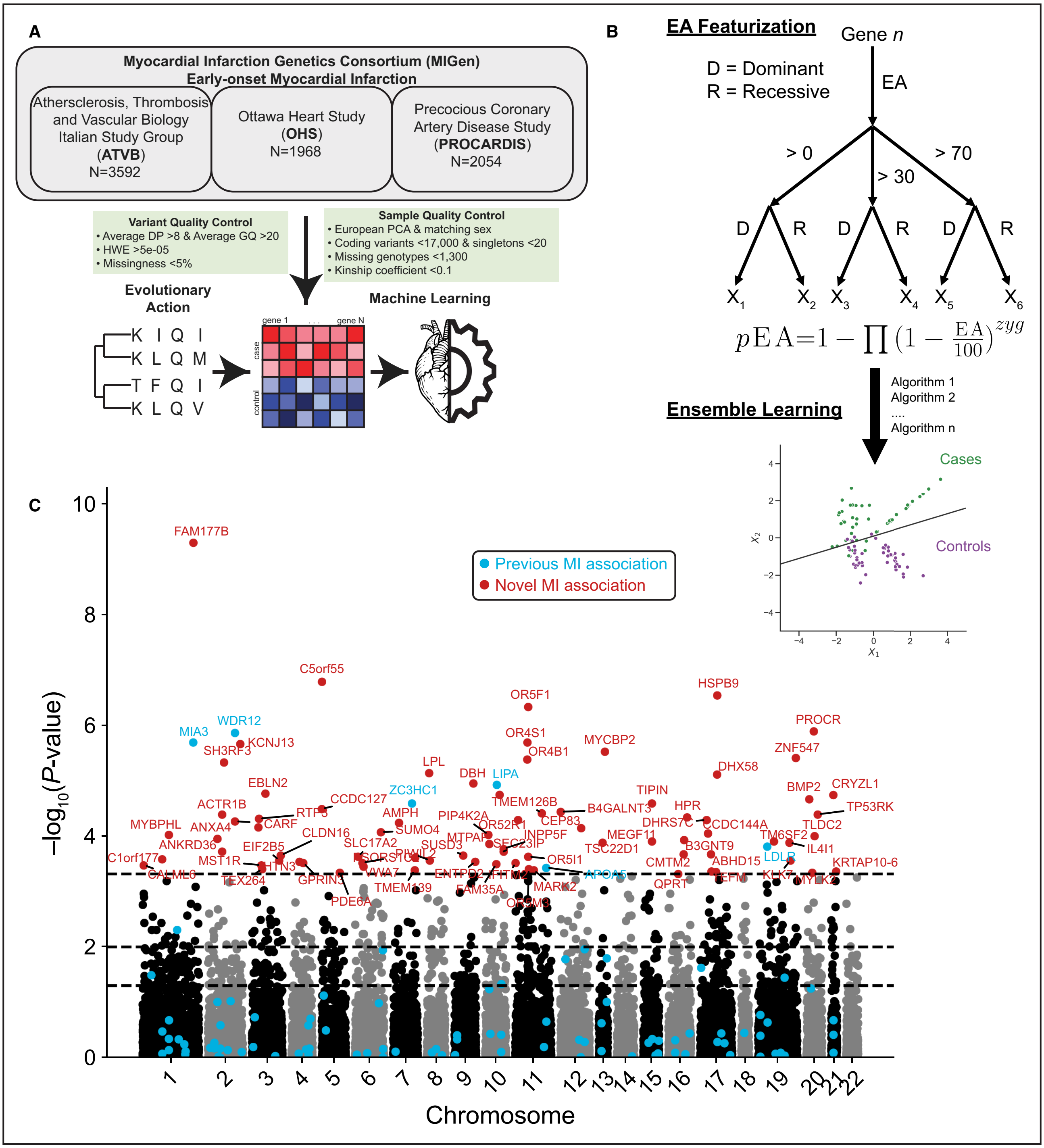


Figure 1. Overview of study design and Evolutionary Action-Machine Learning (EAML) results.
A, Workflow of sample selection, quality control, and EAML analysis. Samples are individuals from 3 separate Myocardial Infarction Genetics studies with whole exome sequencing performed by the Broad Institute. After variant and sample quality control, all single nucleotide variants are given an Evolutionary Action (EA) functional impact score based on the residue's evolutionary importance and amino acid substitution. These scores are used to train an ensemble machine learning model that ranks each gene's disease association. **B**, Schematic of how EAML uses variants and inheritance hypotheses to identify genes associated with early-onset myocardial infarction risk. **C**, Manhattan plot of EAML results. Red dots are genes found by EAML (adjusted *P*<0.1), and blue dots are genes found in previous myocardial infarction genome-wide association studies. DP indicates depth of coverage; GQ, genotype quality; HWE, Hardy-Weinberg Equilibrium; and PCA, principal component analysis.

(P=1.05×10^{-4}), type 2 diabetes (P=1.21×10^{-4}), and antithrombotic agent use (P=3.00×10^{-4}). Notably, *LDLR*, *APOA5*, and *LPL* occurred repeatedly in these enrichments; all of which are extremely important in lipid metabolism and the underlying pathology of cardiovascular disease. Next, we compared the 79 EOMI candidate genes against an aggregated set of reference genes from curated MI associations in GWAS Catalog[67] (n=75) and clinical MI variants in the ClinVar clinical variant database[68] (n=36) (Table S3). The EAML candidates significantly overlapped (n=6) with the union of the 2 reference gene sets (Figure 2B, P=8.55×10^{-6}: *WDR12*, *MIA3, LIPA*, *LDLR*, *APOA5*, and *ZC3HC1*). Because most of the candidates did not overlap with the GWAS Catalog or ClinVar reference gene sets, we evaluated the degree to which our candidate genes colocalized with previously reported CAD GWAS loci. A total of 19 candidates (Table S4) fell within 500 kb of a CAD GWAS Catalog locus (Figure 2C; P=0.0018), showing that our EAML prioritized genes are related to previously identified CAD loci. These data show how the 79 EOMI candidate genes replicate previously reported cardiovascular disease associations and are enriched in known CAD biological processes.

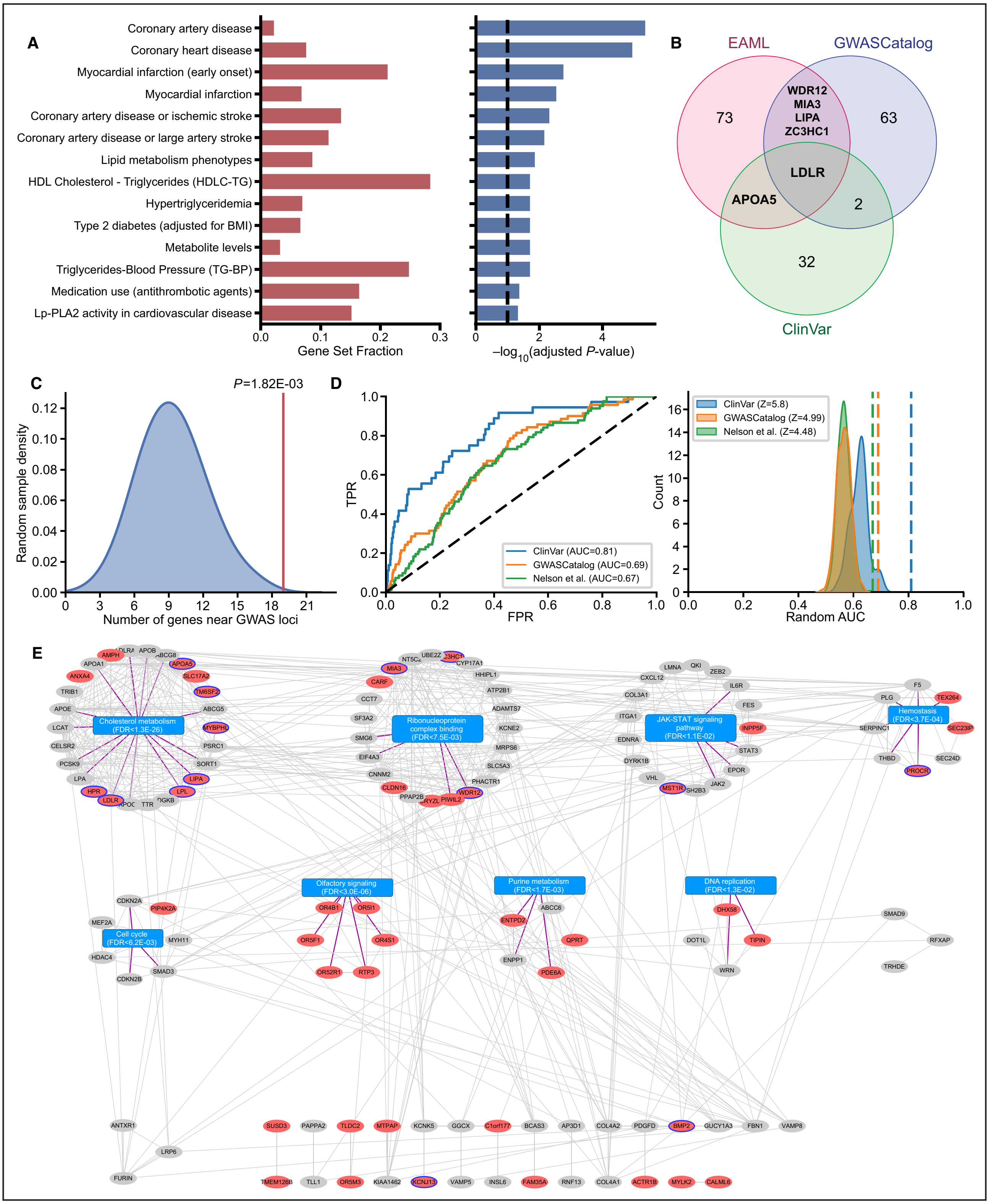
EOMI Candidate Genes Exhibit Significant Network Relatedness and Connectivity to Cardiovascular Risk Genes and Biology

To investigate the mutual interactions of the EOMI candidates, we assessed connectivity of the 79 genes across the STRINGv11 protein–protein interaction network. We found significant connectivity between 23 EOMI candidates (P=1.36×10^{-3}) (Figure S7A), although EAML does not make any a priori assumptions about network connectivity. These genes fall into local network clusters enriched for cholesterol and lipid homeostasis, olfaction, and vascular smooth muscle contraction. Next, to assess the relative network proximity of our 79 EOMI candidate genes to the GWAS Catalog and ClinVar reference gene sets, we used a graph-based information diffusion algorithm (nDiffusion)[62] that measures the closeness of interactions between 2 gene sets in any input network (eg, STRING). The 79 EOMI candidate genes are highly and significantly (Z scores=5.80, 4.99, 4.48) connected to the 2 reference gene sets, as well as a third CAD-associated gene set[11] (Figure 2D, Figure S7B, Table S3). We observed the highest connectivity between EAML genes and ClinVar reference genes (AUC=0.81, Z=5.80). These data show that the 79 EAML genes are significantly connected to known CAD biology in the context of a protein–protein interaction network, implicating potential novel associations and biology.

To further understand how the EOMI candidate genes are related to known cardiovascular risk genes, we built an interaction network between our 79 genes and the GWAS Catalog and ClinVar reference gene sets. The resulting candidate-reference gene hybrid network was highly enriched for protein–protein interactions (P<1.0×10^{-16}) and allowed us to visualize first neighbor connections between the 79 candidates and the reference genes (Figure 2E). We then performed Markov clustering on the hybrid network to identify densely connected regions and subsequently tested for pathway enrichment in the clusters containing at least 1 EOMI candidate gene. Of the 22 Markov clusters, 16 contained at least 1 EOMI candidate. Pathway enrichment of the clusters with g:Profiler[69] identified CAD-relevant biology including lipid metabolism, inflammation, blood coagulation and platelet degranulation, purine and nicotinate/nicotinamide metabolism, and transcriptional regulation (Table S5). The genes related to lipid metabolism fall into the largest cluster, including notable cardiovascular disease risk factors, namely *LDLR*, *APOA5*, *LIPA*, and *LPL*.[13,17,70] Additionally, this cluster contains *HPR* and *ANXA4*, both of which may indirectly impact lipid metabolism through interactions with apolipoprotein L1 (apoL-I)–containing HDL and phospholipids, respectively. It is also noteworthy that *MYBPHL* has been shown to regulate ventricular and atrial conduction and is associated with dilated cardiomyopathy.[71] Other pathways of interest include a cluster with 2 EOMI candidates, *INPP5F* and *MST1R*, linked to inflammation and a cluster related to blood coagulation and platelet degranulation with 3 candidates, *PROCR*, *SEC23IP*, and *TEX264*. The inflammation cluster centers around JAK/STAT signaling, an important inflammatory pathway that is an established target for modulating cardiovascular risk.[72] Additionally, *PROCR* has an essential role in regulating anticoagulation through protein C levels[73] and has previously been associated with CAD and venous thromboembolism.[74,75] These data show that EOMI candidates encompass important biological processes essential to cardiovascular disease through links to known CAD genes.

EOMI Candidates Have Significant Relative Effects on Cardiovascular Disease Risk

To characterize the relative risk associated with each of the 79 candidate EOMI genes, we calculated allelic odds ratios (ORs), aggregating all nonsynonymous variants within each gene. Forty-nine of the 79 EAML candidates were significantly associated with disease status (adjusted P<0.01), of which 15 were associated with increased risk (OR >1), and 34 were associated with decreased risk (OR <1) (Figure 3A). Genes linked to familial hypercholesterolemia like *LDLR* (OR, 1.28) and *APOA5* (OR, 1.2) showed the strongest associations with increased risk, alongside *CALML6* (OR,



1.25). Genes associated with increased risk also included other known CAD risk factors, *WDR12*, *CARF*, *DHX58*, and *LIPA* with OR values of 1.18, 1.15, 1.14, and 1.09, respectively. Surprisingly, most genes were associated with protection from disease, including *MIA3* (OR, 0.87) and *PROCR* (OR, 0.81), despite both genes being known for associations with increased risk of CAD and thromboembolisms, respectively.

Next, to map the risk of each of the 79 EOMI candidates more finely, we tested if specific variants within

Figure 2. Enrichment of Evolutionary Action–Machine Learning (EAML) candidates for direct overlap or interactions with known cardiovascular associations in clinical and genome-wide association studies (GWAS) data.

A, Enrichment for traits with genome-wide associations, performed using the functional mapping and annotation web portal. The left bar plot illustrates the fraction of genes for each enriched trait that overlaps with EAML genes. The right bar plot shows the adjusted *P* value for each enrichment. **B**, Overlap between EAML candidates and myocardial infarction-associated genes from ClinVar and GWAS Catalog databases. **C**, A density plot illustrating enrichment for EAML candidates within 500 kb of established coronary artery disease GWAS loci. The density plot represents the colocalization between GWAS loci and randomly sampled gene sets, and the red line represents the observed colocalization with EAML. The *P* value was calculated using a *Z* test. **D**, Receiver operating characteristic curves for network diffusion from EAML candidates to ClinVar- and GWAS-mapped genes (left). Distributions of areas under the curve (AUCs) based on 100 randomly sampled, degree-matched target gene sets (right). Dashed lines represent the experimental AUCs. **E**, Network-based clustering and functional enrichment of EAML candidates and previously reported myocardial infarction genes. The network was built using 79 EAML candidates and 103 known coronary artery disease risk genes from GWAS Catalog and ClinVar that interacted with one another in STRING version 11 (confidence >0.4). Modules were created using the Markov clustering algorithm with an inflation parameter of 3. Functional enrichment was performed for each cluster using g:Profiler. Red nodes represent EAML genes, and blue borders represent PubMed comentions with cardiovascular disease. BMI indicates body mass index; Lp-PLA2, lipoprotein-associated phospholipase A2; FDR, false discovery rate; FPR, false positive rate; and TPR, true positive rate.

each gene were associated with disease risk. We found 8 variants to be significantly associated with increased risk and 9 variants associated with decreased risk (adjusted $P<0.01$) (Figure 3B). All associations were among common variants with allele frequencies ranging from 0.03 to 0.50 and included variants within genes with previously identified noncoding associations like *MIA3* (K605R, EA=14; E881G, EA=56), *ZC3HC1* (R363H, EA=34), and *DHX58* (Q425R, EA=8). In *MIA3* and *MST1R* (Figure 3C), we identified additional variants with opposing EA scores and identical ORs, potentially due to linkage disequilibrium between the variants. In *FAM177B*, we identified a variant (I3S) that corresponded with increased risk (OR, 1.18) and has recently been associated with CAD in a gene-based meta-analysis along with the substitution Q523R in *MST1R*.[76] These data illustrate that the 79 candidate genes are clinically important by increasing or decreasing the relative risk of EOMI and reveal variants of potentially novel mechanistic interest.

Cardiovascular *cis*-eQTLs Are Enhanced in EOMI Candidate Genes

In addition to being impacted by variation that directly affects protein function, a gene may affect disease risk through regulatory variation. Importantly, regulatory variation can modify the penetrance of protein coding variants.[77,78] To determine if the EOMI candidates contain evidence for *cis*-eQTLs, we queried each candidate gene in the GTEx database[63] for *cis*-eQTL associations within cardiovascular tissues. We found that 67 EOMI candidates show overlapping eQTLs in at least 1 cardiovascular tissue (Figure S8A). Among the candidate genes, neither *APOA5* nor *KCNJ13* showed evidence of significant changes in differentially expressed gene levels, despite their previous strong associations with CAD. However, GTEx contains samples without any specific phenotypes, and these data may not represent how expression is regulated in the context of cardiovascular disease. To identify EOMI candidates containing regulatory associations specific to cardiovascular disease, we searched the eQTL summary statistics from STARNET,[64] a RNAseq-based study of 600 patients with cardiovascular disease that contains 8 291 095 eQTLs mapped to 14 174 genes across 7 cardiovascular tissues. Fifty-three EOMI candidate genes intersected with reported eQTLs in at least 1 of the 7 tissues (Figure S8B). This included EOMI candidate genes that overlapped with the GWAS Catalog and ClinVar reference gene sets, except for *KCNJ13*. Six genes overlap with at least 1 eQTL in all 7 tissues, namely *CCDC127*, *CCDC144A*, *DHX58*, *PIP4K2A*, *TEX264*, and *TIPIN* (Table S6). It is noteworthy that *APOA5* and *LDLR* are linked to eQTLs specific to liver tissue, which play an important role in regulating triglyceride metabolism[79,80] and plasma lipid clearance.[81] These data show that many EOMI candidates are associated with differentially expressed gene levels in cardiovascular tissues, further supporting their importance for CAD progression.

EOMI Candidates Are Enriched for Cardiovascular Effects in Mice

To evaluate whether alterations in EOMI candidates directly drive or modulate biological changes that impact cardiovascular health in animal models, we turned to the Mouse Genome Informatics database.[59] In total, there are 18 122 human genes with mouse orthologs present in the Mouse Genome Informatics database, and 2845 of these exhibit a cardiovascular system phenotype in at least 1 mouse model. Of the 79 EOMI genes tested, 66 were altered in at least 1 mouse model within the database. Of these, we found significant alterations in 19 genes associated with a cardiovascular system phenotype ($P=5.04\times10^{-3}$) when compared with randomly sampled gene sets (average number of cardiovascular phenotypes for random sets=9.3; Z=3.04). Notable overlapped genes are *LDLR*, *LPL*, and *LIPA*, as well as genes with strong common variant associations with CAD, namely *MIA3* and *DHX58*. We also found that

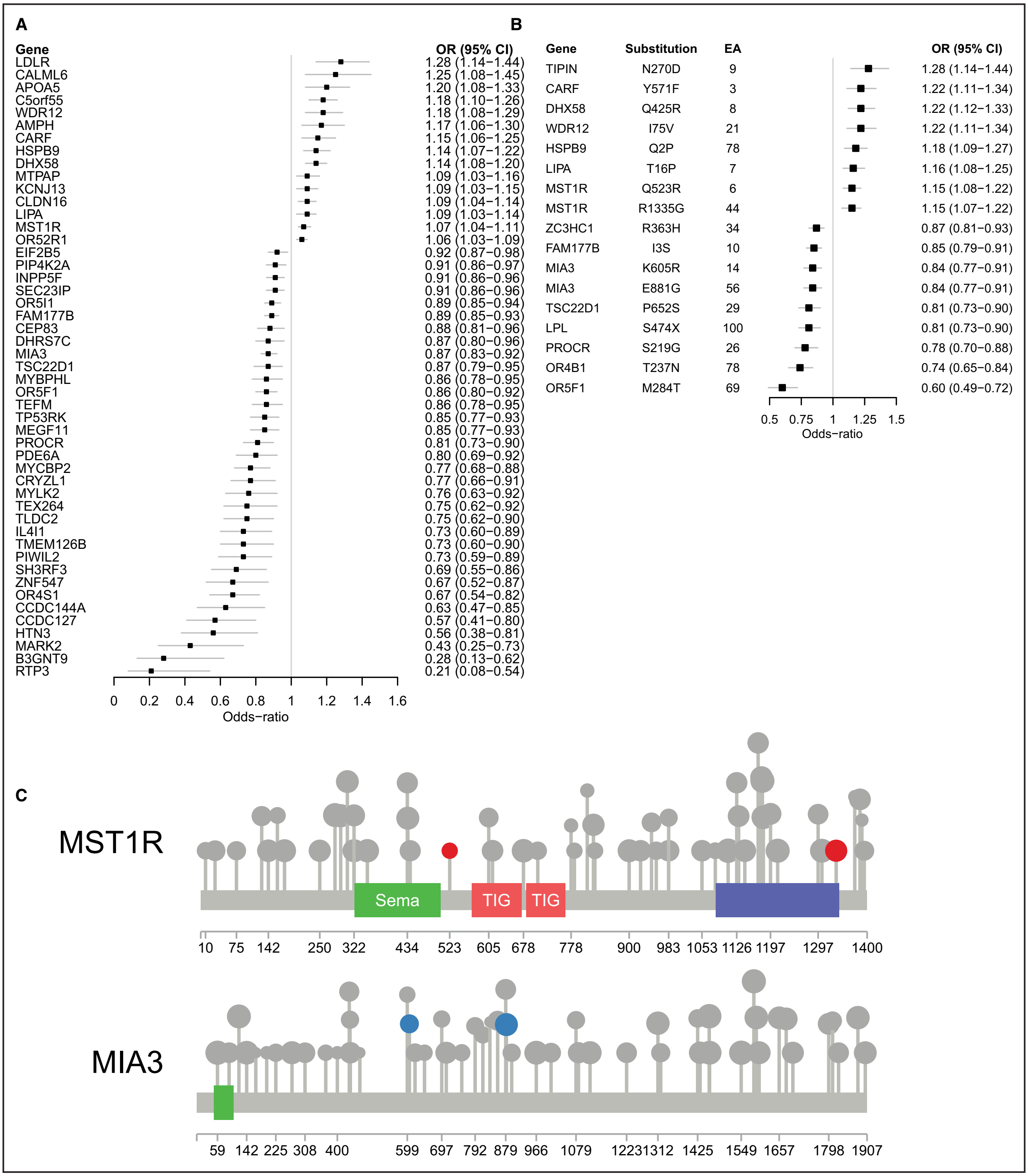


Figure 3. Estimated contributions of Evolutionary Action–Machine Learning (EAML) candidate variants to myocardial infarction (MI) risk.

A, Aggregate odds ratios (ORs) of EAML candidates based on nonsynonymous single nucleotide variants (SNVs) (adjusted *P*<0.01). **B**, ORs of nonsynonymous SNVs with adjusted *P*<0.01 in EAML candidates. **C**, Lollipop plot of nonsingleton missense SNVs in *MST1R* and *MIA3*. Circle size corresponds to Evolutionary Action (EA) score, and colored SNVs are significantly associated with early-onset MI in variant-level OR analysis (adjusted *P*<0.01). Red color corresponds to OR >1, and blue corresponds to OR <1.

4 genes without previously reported MI associations (*BMP2*, *TEFM*, *MARK2*, and *DBH)* affected heart physiology and inflammatory response. Despite having no direct CAD associations, *BMP2* has been shown to be essential for proper cardiac development.[82–84] *MST1R* models exhibit increased acute inflammation, whereas models targeting *LEXM*, *B4GALNT3*, and *MARK2* exhibit abnormal T-cell morphology. These data show the phenotypic relevance of the EOMI genes in the context of a model organism, supporting their predicted importance in cardiovascular disease.

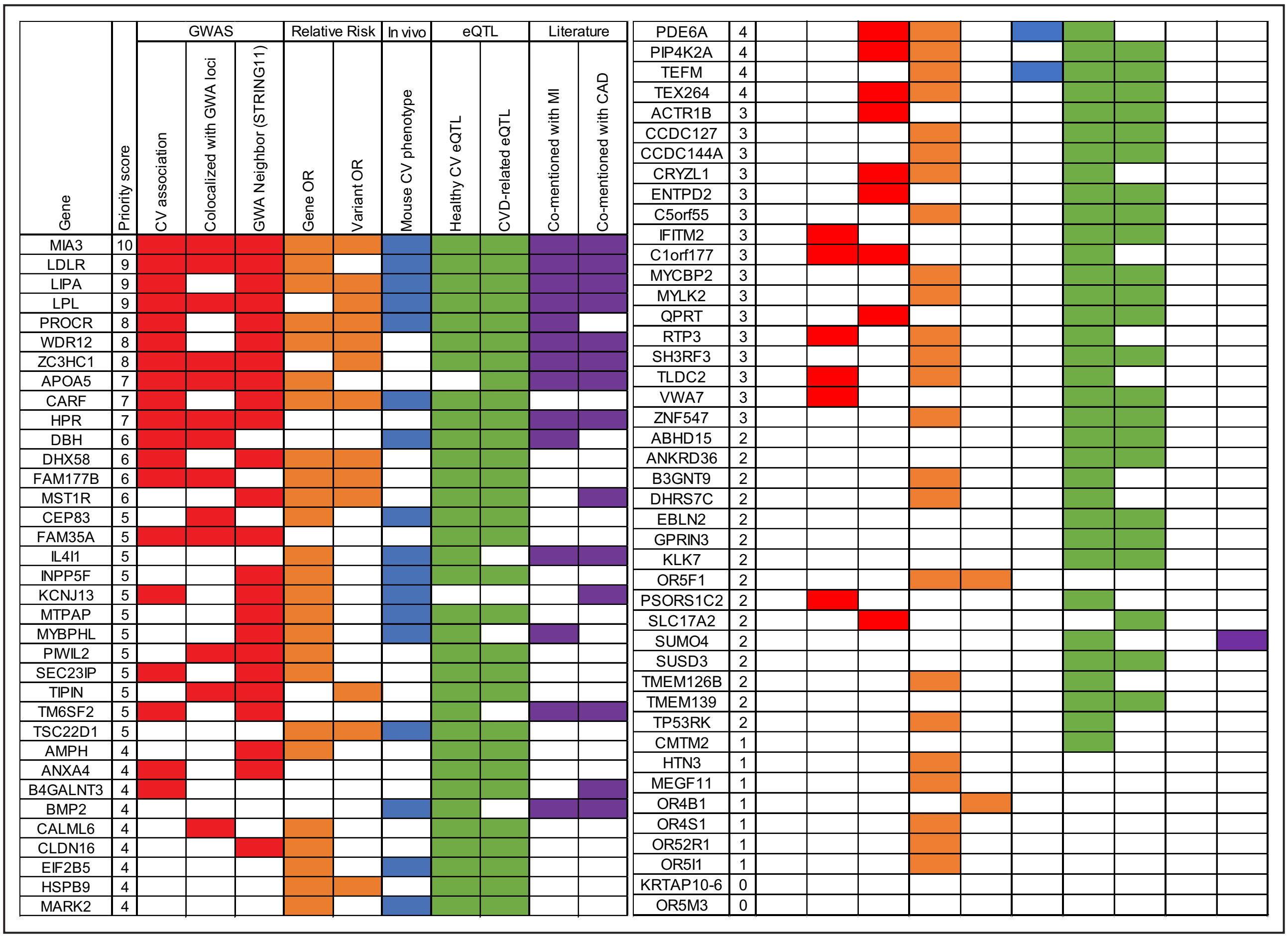
EAML Prioritizes Genes of Biological Importance in Cardiovascular Disease

To prioritize EOMI candidate genes that are most likely to hold biological significance in cardiovascular disease risk and mechanism, we combined the results from our validation analyses to build a gene prioritization score (Figure 4)[85,86] based on 10 equally weighted experiments in 5 overall categories: GWAS, relative risk, in vivo, eQTL, and literature. For the GWAS category, we aggregated criteria related to known cardiovascular associations, colocalization with an established GWAS loci (Figure 2C), and direct interaction with a GWAS-reported gene in STRING (Figure 2E). For the relative risk category, positive criteria include having a statistically significant allelic OR (adjusted P<0.01) or containing any variant with a significant OR (Figure 3). The in vivo criteria include whether the gene showed cardiovascular phenotype evidence in the Mouse Genome Informatics database. The eQTL category contains criteria for the presence of cardiovascular *cis*-eQTLs in GTEx or STARNET (Figure S7). Lastly, the literature category contains evidence of comentions with either CAD or MI in PubMed titles and abstracts. Based on these criteria, *MIA3* was the top gene with positive evidence in every category, followed by 13 genes with priority scores between 6 and 9: *LDLR*, *LIPA*, *LPL*, *PROCR*, *WDR12*, *ZC3HC1*, *APOA5*, *CARF*, *HPR*, *DBH*, *DHX58*, *FAM177B*, and *MST1R*. The top 8 genes all contain established links to CAD biology, whereas the subsequent 6 genes possess mostly circumstantial evidence. Among the potentially novel gene findings with little to no prior literature association with CAD or MI, the highest prioritized candidates were *CARF* and *HPR*, with a score of 7. *CARF* lies within the same region as *WDR12*, another previously reported CAD gene. Both genes are targeted by *cis*-eQTLs within the same locus[87] and contain variants (rs72932557, rs35212307) that are associated with similar levels of increased MI risk. *CARF* was also identified as a key driver in a gene regulatory network derived from the STARNET cohort.[64] *HPR* has functional evidence linking it to cardiovascular biology through hemoglobin and HDL.[88] Next, *DBH*, *DHX58*, *FAM177B*, and *MST1R* had gene prioritization scores of 6. *DBH* is essential to noradrenaline production, with liver-specific eQTLs (Table S6) and knockout mouse models showing defects related to general development, heart morphology, and circulating hormone levels.[89] *DHX58* has evidence from 2 previous CAD association studies, 1 of which identifies the same risk-increasing coding variant as our study (rs2074158; OR, 1.22; P=9.58×10^{-6}).[70,90] *FAM177B* was scored as the top EAML candidate gene, yet there are no direct studies linking it to MI or CAD. However, our analysis suggests it is associated with a protective effect (OR, 0.89 [95% CI, 0.85–0.93]; P=2.×10^{-5}), and intergenic variants have been associated with coronary artery bypass grafting.[91] Another interesting and novel candidate with a gene prioritization score of 5 is *INPP5F*. It has not previously been associated with cardiovascular phenotypes and fails to colocalize with established risk loci, yet we find evidence supporting an association with CAD. The *INPP5F* gene-level (OR, 0.91 [95% CI, 0.86–0.96]; P=6.2×10^{-4}) risk appears mostly driven by the protective variant rs318805 (OR, 0.88 [95% CI, 0.0.82–0.94]; P=3.3×10^{-4}), which aligns with the functional role it plays in inhibiting *STAT3* and its antiapoptotic/proangiogenic activity.[92,93] These data show that by aggregating different types supporting evidence, we reaffirm genes with known CAD associations and prioritize genes with novel clinical and functional insights.

DISCUSSION

Previously, we presented a novel genomic analysis framework, EAML, that combines ensemble machine learning with a continuous functional impact score for coding variants for discovering genotype–phenotype associations.[33] Here, for cardiovascular disease and EOMI, we identified 79 genes with underlying mutational patterns specific to EOMI, of which 60 were not previously linked to CAD, to the best of our knowledge.

Consistent with past works, our study shows that lipid metabolism continues to be the pathway most strongly associated with CAD risk. Mutations in *LDLR*, *LPL*, and *APOA5* are known for their direct links to lipid metabolism and trafficking, with rare variants often being causal for familial hyperlipidemias.[94–96] Although no individual variants in *LDLR* or *APOA5* in the MIGen cohort are significantly associated with MI status, presence of any mutation in either of the genes shows an association with increased risk (ORs: *LDLR* 1.28 and *APOA5* 1.20). In addition to these established risk genes, we identified *HPR*, *MYBPHL*, *AMPH*, *ANXA4*, and *SLC17A2*, all of which have little to no direct evidence linking them to CAD. In particular, *HPR* was ranked in the top 10 genes in the prioritization table, has appeared in multiple studies associated with both



Gene	Priority score	GWAS			Relative Risk		In vivo	eQTL		Literature	
		CV association	Colocalized with GWA loci	GWA Neighbor (STRING11)	Gene OR	Variant OR	Mouse CV phenotype	Healthy CV eQTL	CVD-related eQTL	Co-mentioned with MI	Co-mentioned with CAD
MIA3	10	🟥	🟥	🟥	🟧	🟧	🟦	🟩	🟩	🟪	🟪
LDLR	9	🟥	🟥	🟥	🟧		🟦	🟩	🟩	🟪	🟪
LIPA	9	🟥		🟥	🟧	🟧	🟦	🟩	🟩	🟪	🟪
LPL	9	🟥	🟥	🟥		🟧	🟦	🟩	🟩	🟪	🟪
PROCR	8	🟥		🟥	🟧	🟧	🟦	🟩	🟩	🟪	
WDR12	8	🟥		🟥	🟧	🟧		🟩	🟩	🟪	🟪
ZC3HC1	8	🟥	🟥	🟥		🟧		🟩	🟩	🟪	🟪
APOA5	7	🟥	🟥	🟥	🟧				🟩	🟪	🟪
CARF	7	🟥		🟥	🟧	🟧	🟦	🟩	🟩		
HPR	7	🟥	🟥	🟥				🟩	🟩	🟪	🟪
DBH	6	🟥	🟥				🟦	🟩	🟩	🟪	
DHX58	6	🟥		🟥	🟧	🟧		🟩	🟩		
FAM177B	6	🟥	🟥		🟧	🟧		🟩	🟩		
MST1R	6			🟥	🟧	🟧		🟩	🟩		🟪
CEP83	5		🟥		🟧		🟦	🟩	🟩		
FAM35A	5	🟥	🟥	🟥				🟩	🟩		
IL4I1	5				🟧		🟦	🟩		🟪	🟪
INPP5F	5			🟥	🟧		🟦	🟩	🟩		
KCNJ13	5	🟥		🟥	🟧		🟦				🟪
MTPAP	5			🟥	🟧		🟦	🟩	🟩		
MYBPHL	5			🟥	🟧		🟦	🟩		🟪	
PIWIL2	5		🟥	🟥	🟧			🟩	🟩		
SEC23IP	5	🟥		🟥	🟧			🟩	🟩		
TIPIN	5		🟥	🟥		🟧		🟩	🟩		
TM6SF2	5	🟥		🟥				🟩		🟪	🟪
TSC22D1	5				🟧	🟧	🟦	🟩	🟩		
AMPH	4			🟥	🟧			🟩	🟩		
ANXA4	4	🟥		🟥				🟩	🟩		
B4GALNT3	4	🟥						🟩	🟩		🟪
BMP2	4						🟦	🟩		🟪	🟪
CALML6	4		🟥		🟧			🟩	🟩		
CLDN16	4			🟥	🟧			🟩	🟩		
EIF2B5	4				🟧		🟦	🟩	🟩		
HSPB9	4				🟧	🟧		🟩	🟩		
MARK2	4				🟧		🟦	🟩	🟩		
PDE6A	4			🟥	🟧		🟦	🟩			
PIP4K2A	4			🟥	🟧			🟩	🟩		
TEFM	4				🟧		🟦	🟩	🟩		
TEX264	4			🟥	🟧			🟩	🟩		
ACTR1B	3			🟥				🟩	🟩		
CCDC127	3				🟧			🟩	🟩		
CCDC144A	3				🟧			🟩	🟩		
CRYZL1	3			🟥	🟧			🟩			
ENTPD2	3			🟥				🟩	🟩		
C5orf55	3				🟧			🟩	🟩		
IFITM2	3		🟥					🟩	🟩		
C1orf177	3		🟥	🟥				🟩			
MYCBP2	3				🟧			🟩	🟩		
MYLK2	3				🟧			🟩	🟩		
QPRT	3			🟥				🟩	🟩		
RTP3	3		🟥		🟧			🟩			
SH3RF3	3				🟧			🟩	🟩		
TLDC2	3		🟥		🟧			🟩			
VWA7	3		🟥					🟩	🟩		
ZNF547	3				🟧			🟩	🟩		
ABHD15	2							🟩	🟩		
ANKRD36	2							🟩	🟩		
B3GNT9	2				🟧			🟩			
DHRS7C	2				🟧			🟩			
EBLN2	2							🟩	🟩		
GPRIN3	2							🟩	🟩		
KLK7	2							🟩	🟩		
OR5F1	2				🟧	🟧					
PSORS1C2	2		🟥					🟩			
SLC17A2	2			🟥					🟩		
SUMO4	2							🟩			🟪
SUSD3	2							🟩	🟩		
TMEM126B	2				🟧			🟩			
TMEM139	2							🟩	🟩		
TP53RK	2				🟧			🟩			
CMTM2	1							🟩			
HTN3	1				🟧						
MEGF11	1				🟧						
OR4B1	1					🟧					
OR4S1	1				🟧						
OR52R1	1				🟧						
OR5I1	1				🟧						
KRTAP10-6	0										
OR5M3	0										

Figure 4. Prioritization of Evolutionary Action–Machine Learning candidates.
Criteria include (1) mapped to previously reported cardiovascular (CV) genome-wide association studies (GWAS), (2) located within 500 kb of a previously reported GWAS locus, (3) first-neighbor interaction with a mapped GWAS gene in the STRING protein-protein interaction network, (4) gene-based odds ratio (OR) with adjusted *P*<0.01, (5) at least 1 variant with OR with adjusted *P*<0.01, (6) associated with CV phenotype in Mouse Genome Informatics database, (7) reported expression quantitative trait locus (eQTL) association in CV healthy tissue in the Genotype-Tissue Expression database, (8) reported eQTL association in CV tissues in the STARNET (Stockholm-Tartu Atherosclerosis Reverse Network Engineering Task) study, (9) text or abstract comention with myocardial infarction (MI), and (10) text or abstract comention with coronary artery disease (CAD). Each category is equally weighted, and the priority score is the sum of all categories. Colored fields indicate positive evidence for the given gene. CVD indicates cardiovascular disease; and GWA, genome-wide association.

low-density lipoprotein[97] and total cholesterol levels,[98] and is expressed exclusively in the heart and liver in GTEx samples. The *HPR* protein is known for its role in innate immune protection against trypanosomes through association with apolipoprotein L1–containing HDL particles,[99] yet our results suggest that its role may extend beyond that in terms of impact on CAD risk. Additionally, *MYBPHL* has previously been characterized for its function in cardiac conduction through atrial cardiomyocytes.[100] There is also strong evidence associating the *MYBPHL*-containing locus 1p13.3 with decreased low-density lipoprotein cholesterol,[101] which aligns with our data that *MYBPHL* mutations are associated with a protective effect on EOMI risk (Figure 3A). Although it is well known that lipid metabolism is important in cardiovascular health, EAML further supports this by placing multiple associated genes in a functional context.

Following lipid metabolism, inflammation is the other major mechanism involved in all stages of atherosclerotic progression, from plaque formation[102,103] to post-MI recovery.[104] Three EAML genes, *INPP5F*, *MST1R*, and *DHX58*, were linked to inflammatory functions. First, *INPP5F* has a gene prioritization score of 5 and lacks previous CAD evidence (Figure 4), yet showed an allelic association with EOMI risk and was directly linked to *STAT3* (Figure 2E),[92] an important inflammatory signaling regulator. *INPP5F* encodes the *SAC2* protein, an inositol 4-phospatase involved in the endocytic recycling pathway.[105,106] *Inpp5f*$^{-/-}$ mice exhibit increased susceptibility to stress-induced cardiac hypertrophy, and cardiac-specific overexpression led to a decreased hypertrophic response.[107] Through its inhibition of *STAT3*, *INPP5F* may play an important role in JAK/STAT immune signaling, which is activated in response to acute MI.[72] When acutely activated, this key inflammatory pathway invokes a cytokine cascade that exerts a cardioprotective effect through the regulation of myocyte survival, whereas chronic activation of JAK/STAT signaling leads to cardiac remodeling and a decline in heart function.[108] *INPP5F* likely plays a poorly studied role in these processes. A second potentially novel risk gene, *MST1R* (scored at 6), is a receptor tyrosine kinase that regulates wound healing and plays a role in both chronic and acute inflammation through macrophage recruitment.[109,110] Our analyses showed that *MST1R* is connected in STRING to *EPOR*, a regulator of *JAK2* (Figure 2E), and contains a single nucleotide polymorphism that is directly associated with increased EOMI risk (R1335G; OR, 1.15; EA=43.52) (Figure 3B). *MST1R* has recently been identified by gene-based association tests for CAD,[76] and *MST1R* null mice have impaired nitric oxide levels and increased tissue damage in response to acute stress.[111] Third, *DHX58* represents a previously reported yet understudied association in the context of CAD.[70,90] Although it does not link to the same immune pathways, evidence has shown that the downstream MAVS (mitochondrial antiviral signaling protein) regulates inflammation and fibrosis through NF-κB and MAPK with reduced expression of *MAVS* being associated with improved cardiac function.[112] When linked to the increased relative risk from *DHX58* coding mutants in our own study (Figure 3), this suggests that *DHX58* has direct functional importance in CAD.

Our findings also highlighted genes related to other mechanisms important to cardiovascular health, including blood clotting, purine metabolism, and nicotinamide adenine dinucleotide (NAD) metabolism. EAML genes linked to these pathways include *PROCR*, *SEC23IP*, *TEX264*, *PDE6A*, *ENTPD2*, and *QPRT*. The blood clotting cluster centers around *PROCR*, which contains a variant we report as protective (rs867186; OR, 0.78), matching what was previously reported in a GWAS within the ATVB cohort.[113] This same variant has also been previously associated with increased levels of protein C,[114] increased levels of factor VII,[115] increased risk of venous thrombosis,[116] and decreased risk of CAD.[74] Also included in the blood clotting cluster are *SEC23IP* and *TEX264*, both of which are novel risk genes that appear to be involved in endoplasmic reticulum–related functions. *SEC23IP* is involved in coat protein complex II (COPII)-mediated endoplasmic reticulum-to-Golgi trafficking,[117] which is essential for the secretion of clotting-related factors, whereas *TEX264* is involved in endoplasmic reticulum-phagy in response to nutrient stress.[118,119] *SEC23IP* has also previously been associated with type 2 diabetes[120] and HDL cholesterol levels.[121] A final interesting cluster is one primarily enriched for purine and NAD metabolism, containing 3 novel EAML candidates: *PDE6A*, *ENTPD2*, and *QPRT*. Although both pathways are less studied in the context of cardiovascular health, uric acid (a product of purine metabolism) has been associated with long-term cardiovascular risk,[122,123] and experimental evidence has shown that NAD+ elevation can protect against cardiovascular outcomes in preclinical models.[124,125]

Although the EAML method shows many advantages such as the incorporation of a gene-based impact score derived from evolutionary history and the use of ensemble machine learning, we could not consider hyperparameter optimization for each individual classifier. Due to the algorithmic complexity of training 9 classifiers for 18 000 genes, optimization would lead to an exponential increase in computing time and required computing power, without a guaranteed performance improvement. To address this, we performed 10-fold cross-validation for each individual classifier to minimize bias and overfitting/underfitting of the models by repeatedly testing each gene and averaging the results, and we strengthened our confidence in the resulting candidate genes through multiple independent validation experiments. Specifically, we illustrate that these genes are reliably linked to CAD and have performed an extensive review of published evidence, showing that many of these genes have biologically relevant functions. Therefore, our observations suggest that hyperparameter tuning was not a significant issue in the identification of CAD-associated genetic risk factors using EAML in the current study.

In conclusion, this study provides new insights into cardiovascular genetics by extending exome-wide associations with the combination of evolutionary information and supervised machine learning. Although EAML is still limited to analyzing individual genetic risk factors as in standard GWAS methods, it is broadly applicable to case–control whole exome studies. In addition to solidifying the role of several common genetic risk factors, EAML discovered both novel and previously known risk genes that had only been associated with CAD through noncoding variation. Furthermore, several EAML candidates without direct cardiovascular associations were closely related to established CAD risk loci, and linked together in gene clusters enriched in biological functions related to known disease mechanisms. Lastly, our study illustrates that protein-coding variation has a significant impact on complex disease risk. Our findings have the added benefit of being directly targetable in future mechanistic studies and are applicable to polygenic risk methods, which can further inform cardiovascular causes.

ARTICLE INFORMATION

Received December 5, 2022; accepted July 11, 2023.

Affiliations

Department of Molecular and Human Genetics (D.S., K.L., J.A., T.B., O.L.), and Computational & Integrative Biomedical Research Center (O.L.), Baylor College of Medicine, Houston, TX.

Acknowledgments

The authors thank the staff and participants of the ATVB study, OHS, and PROCARDIS study for their important contributions.

Sources of Funding

This work was supported by the National Institutes of Health 1R01AG074009, 1U01AG068214, and 5R 01AG061105.

Disclosures

None.

Supplemental Material

Data S1–S2
Tables S1–S6
Figures S1–S8

REFERENCES

1. Benjamin EJ, Blaha MJ, Chiuve SE, Cushman M, Das SR, Deo R, de Ferranti SD, Floyd J, Fornage M, Gillespie C, et al. Heart disease and stroke statistics—2017 update: a report from the American Heart Association. *Circulation*. 2017;135:e146–e603. doi: 10.1161/CIR.0000000000000485
2. Virani SS, Alonso A, Aparicio HJ, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Cheng S, Delling FN, et al. Heart disease and stroke statistics—2021 update: a report From the American Heart Association. *Circulation*. 2021;143:E254–E743. doi: 10.1161/CIR.0000000000000950
3. Zdravkovic S, Wienke A, Pedersen NL, Marenberg ME, Yashin AI, De Faire U. Heritability of death from coronary heart disease: a 36-year follow-up of 20 966 Swedish twins. *J Intern Med*. 2002;252:247–254. doi: 10.1046/j.1365-2796.2002.01029.x
4. Won H-H, Natarajan P, Dobbyn A, Jordan DM, Roussos P, Lage K, Raychaudhuri S, Stahl E, Do R. Disproportionate contributions of select genomic compartments and cell types to genetic risk for coronary artery disease. *PLoS Genet*. 2015;11:e1005622. doi: 10.1371/journal.pgen.1005622
5. Khera AV, Kathiresan S. Genetics of coronary artery disease: discovery, biology and clinical translation. *Nat Rev Genet*. 2017;18:331–344. doi: 10.1038/nrg.2016.160
6. Hartmann K, Seweryn M, Sadee W. Interpreting coronary artery disease GWAS results: a functional genomics approach assessing biological significance. *PLoS One*. 2022;17:e0244904. doi: 10.1371/journal.pone.0244904
7. Clarke SL, Assimes TL. Genome-wide association studies of coronary artery disease: recent progress and challenges ahead. *Curr Atheroscler Rep*. 2018;20:47. doi: 10.1007/s11883-018-0748-4
8. Preuss M, König IR, Thompson JR, Erdmann J, Absher D, Assimes TL, Blankenberg S, Boerwinkle E, Chen L, Cupples LA, et al. Design of the Coronary ARtery DIsease genome-wide replication and meta-analysis (CARDIoGRAM) study: a genome-wide association meta-analysis involving more than 22 000 cases and 60 000 controls. *Circ Cardiovasc Genet*. 2010;3:475–483. doi: 10.1161/CIRCGENETICS.109.899443
9. Deloukas P, Kanoni S, Willenborg C, Farrall M, Assimes TL, Thompson JR, Ingelsson E, Saleheen D, Erdmann J, Goldstein BA, et al. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet*. 2013;45:25–33. doi: 10.1038/ng.2480
10. Wang Q, Dhindsa RS, Carss K, Harper AR, Nag A, Tachmazidou I, Vitsios D, Deevi SVV, Mackay A, Muthas D, et al. Rare variant contribution to human disease in 281,104 UK Biobank exomes. *Nature*. 2021;597:527–532. doi: 10.1038/s41586-021-03855-y
11. Nelson CP, Goel A, Butterworth AS, Kanoni S, Webb TR, Marouli E, Zeng L, Ntalla I, Lai FY, Hopewell JC, et al. Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nat Genet*. 2017;49:1385–1391. doi: 10.1038/ng.3913
12. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, Taliun SAG, Corvelo A, Gogarten SM, Kang HM, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed program. *Nature*. 2021;590:290–299. doi: 10.1038/s41586-021-03205-y
13. Do R, Stitziel NO, Won H-H, Jørgensen AB, Duga S, Angelica Merlini P, Kiezun A, Farrall M, Goel A, Zuk O, et al. Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature*. 2015;518:102–106. doi: 10.1038/nature13917
14. Zheng Q, Zhang Y, Jiang J, Jia J, Fan F, Gong Y, Wang Z, Shi Q, Chen D, Huo Y. Exome-wide association study reveals several susceptibility genes and pathways associated with acute coronary syndromes in Han Chinese. *Front Genet*. 2020;11:336. doi: 10.3389/fgene.2020.00336
15. Park J, Lucas AM, Zhang X, Chaudhary K, Cho JH, Nadkarni G, Dobbyn A, Chittoor G, Josyula NS, Katz N, et al. Exome-wide evaluation of rare coding variants using electronic health records identifies new gene–phenotype associations. *Nat Med*. 2021;27:66–72. doi: 10.1038/s41591-020-1133-8
16. Abifadel M, Varret M, Rabès J-P, Allard D, Ouguerram K, Devillers M, Cruaud C, Benjannet S, Wickham L, Erlich D, et al. Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. *Nat Genet*. 2003;34:154–156. doi: 10.1038/ng1161
17. Garcia CK, Wilund K, Arca M, Zuliani G, Fellin R, Maioli M, Calandra S, Bertolini S, Cossu F, Grishin N, et al. Autosomal recessive hypercholesterolemia caused by mutations in a putative LDL receptor adaptor protein. *Science*. 2001;292:1394–1398. doi: 10.1126/science.1060458
18. Górski B, Kubalska J, Naruszewicz M, Lubiński J. LDL-R and Apo-B-100 gene mutations in Polish familial hypercholesterolemias. *Hum Genet*. 1998;102:562–565. doi: 10.1007/s004390050740
19. Humphries SE, Whittall RA, Hubbart CS, Maplebeck S, Cooper JA, Soutar AK, Naoumova R, Thompson GR, Seed M, Durrington PN, et al. Genetic causes of familial hypercholesterolaemia in patients in the UK: relation to plasma lipid levels and coronary heart disease risk. *J Med Genet*. 2006;43:943–949. doi: 10.1136/jmg.2006.038356
20. Krogh HW, Mundal L, Holven KB, Retterstøl K. Patients with familial hypercholesterolaemia are characterized by presence of cardiovascular disease at the time of death. *Eur Heart J*. 2016;37:1398–1405. doi: 10.1093/eurheartj/ehv602
21. Agarwala V, Flannick J, Sunyaev S, Altshuler D. Evaluating empirical bounds on complex disease genetic architecture. *Nat Genet*. 2013;45:1418–1427. doi: 10.1038/ng.2804
22. Nikpay M, Stewart AFR, McPherson R. Partitioning the heritability of coronary artery disease highlights the importance of immune-mediated processes and epigenetic sites associated with transcriptional activity. *Cardiovasc Res*. 2017;113:973–983. doi: 10.1093/cvr/cvx019
23. Gazal S, Loh PR, Finucane HK, Ganna A, Schoech A, Sunyaev S, Price AL. Functional architecture of low-frequency variants highlights strength of negative selection across coding and non-coding annotations. *Nat Genet*. 2018;50:1600–1607. doi: 10.1038/s41588-018-0231-8
24. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: genetic interactions create phantom heritability. *Proc Natl Acad Sci*. 2012;109:1193–1198. doi: 10.1073/pnas.1119675109
25. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Christiani DC, Wurfel MM, Lin X. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet*. 2012;91:224–237. doi: 10.1016/j.ajhg.2012.06.007
26. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*. 2012;13:762–775. doi: 10.1093/biostatistics/kxs014
27. Mbatchou J, Barnard L, Backman J, Marcketta A, Kosmicki JA, Ziyatdinov A, Benner C, O'Dushlaine C, Barber M, Boutkov B, et al. Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet*. 2021;53:1097–1103. doi: 10.1038/s41588-021-00870-7
28. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, De Bakker PIW, Daly MJ, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81:559–575. doi: 10.1086/519795
29. Guindo-Martínez M, Amela R, Bonàs-Guarch S, Puiggròs M, Salvoro C, Miguel-Escalada I, Carey CE, Cole JB, Rüeger S, Atkinson E, et

al. The impact of non-additive genetic associations on age-related complex diseases. *Nat Commun*. 2021;12:2436. doi: 10.1038/s41467-021-21952-4

30. Salanti G, Southam L, Altshuler D, Ardlie K, Barroso I, Boehnke M, Cornelis MC, Frayling TM, Grallert H, Grarup N, et al. Underlying genetic models of inheritance in established type 2 diabetes associations. *Am J Epidemiol*. 2009;170:537–545. doi: 10.1093/aje/kwp145
31. Lenz TL, Deutsch AJ, Han B, Hu X, Okada Y, Eyre S, Knapp M, Zhernakova A, Huizinga TWJ, Abecasis G, et al. Widespread non-additive and interaction effects within HLA loci modulate the risk of autoimmune diseases. *Nat Genet*. 2015;47:1085–1090. doi: 10.1038/ng.3379
32. Parvandeh S, Donehower LA, Panagiotis K, Hsu T-K, Asmussen JK, Lee K, Lichtarge O. EPIMUTESTR: a nearest neighbor machine learning approach to predict cancer driver genes from the evolutionary action of coding variants. *Nucleic Acids Res*. 2022;50:e70. doi: 10.1093/nar/gkac215
33. Bourquard T, Lee K, Al-Ramahi I, Pham M, Shapiro D, Lagisetty Y, Soleimani S, Mota S, Wilhelm K, Samieinasab M, et al. Functional variants identify sex-specific genes and pathways in Alzheimer's disease. *Nat Commun*. 2023;14:2765. doi: 10.1038/s41467-023-38374-z
34. Katsonis P, Lichtarge O. A formal perturbation equation between genotype and phenotype determines the evolutionary action of protein-coding variations on fitness. *Genome Res*. 2014;24:2050–2058. doi: 10.1101/gr.176214.114
35. Lichtarge O, Bourne HR, Cohen FE. An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol*. 1996;257:342–358. doi: 10.1006/jmbi.1996.0167
36. Katsonis P, Lichtarge O. CAGI5: objective performance assessments of predictions based on the evolutionary action equation. *Hum Mutat*. 2019;40:1436–1454. doi: 10.1002/humu.23873
37. Kim YW, Al-Ramahi I, Koire A, Wilson SJ, Konecki DM, Mota S, Soleimani S, Botas J, Lichtarge O. Harnessing the paradoxical phenotypes of APOE ε2 and APOE ε4 to identify genetic modifiers in Alzheimer's disease. *Alzheimers Dement*. 2021;17:831–846. doi: 10.1002/alz.12240
38. Lagisetty Y, Bourquard T, Al-Ramahi I, Mangleburg CG, Mota S, Soleimani S, Shulman JM, Botas J, Lee K, Lichtarge O. Identification of risk genes for Alzheimer's disease by gene embedding. *Cell Genomics*. 2022;2:100162. doi: 10.1016/j.xgen.2022.100162
39. Ally A, Balasundaram M, Carlsen R, Chuah E, Clarke A, Dhalla N, Holt RA, Jones SJM, Lee D, Ma Y, et al. Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell*. 2017;169:1327–1341.e23. doi: 10.1016/j.cell.2017.05.046
40. Clarke CN, Katsonis P, Hsu T-K, Koire AM, Silva-Figueroa A, Christakis I, Williams MD, Kutahyalioglu M, Kwatampora L, Xi Y, et al. Comprehensive genomic characterization of parathyroid cancer identifies novel candidate driver mutations and core pathways. *J Endocr Soc*. 2019;3:544–559. doi: 10.1210/js.2018-00043
41. Hsu T-K, Asmussen J, Koire A, Choi B-K, Gadhikar MA, Huh E, Lin C-H, Konecki DM, Kim YW, Pickering CR, et al. A general calculus of fitness landscapes finds genes under selection in cancers. *Genome Res*. 2022;32:916–929. doi: 10.1101/gr.275811.121
42. Koire A, Katsonis P, Kim YW, Buchovecky C, Wilson SJ, Lichtarge O. A method to delineate de novo missense variants across pathways prioritizes genes linked to autism. *Sci Transl Med*. 2021;13:eabc1739. doi: 10.1126/scitranslmed.abc1739
43. Marciano DC, Wang C, Hsu T-K, Bourquard T, Atri B, Nehring RB, Abel NS, Bowling EA, Chen TJ, Lurie PD, et al. Evolutionary action of mutations reveals antimicrobial resistance genes in *Escherichia coli*. *Nat Commun*. 2022;13:3189. doi: 10.1038/s41467-022-30889-1
44. Mannucci PM, Merlini PA, Ardissino D, Barzuini C, Bernardi F, Bernardinelli L, Cavallini C, Celli P, Corsini G, Ferrario M, et al. No evidence of association between prothrombotic gene polymorphisms and the development of acute myocardial infarction at a young age. *Circulation*. 2003;107:1117–1122. doi: 10.1161/01.cir.0000051465.94572.d0
45. McPherson R, Pertsemlidis A, Kavaslar N, Stewart A, Roberts R, Cox DR, Hinds DA, Pennacchio LA, Tybjaerg-Hansen A, Folsom AR, et al. A common allele on chromosome 9 associated with coronary heart disease. *Science*. 2007;316:1488–1491. doi: 10.1126/science.1142447
46. Clarke R, Peden JF, Hopewell JC, Kyriakou T, Goel A, Heath SC, Parish S, Barlera S, Franzosi MG, Rust S, et al. Genetic variants associated with Lp(a) lipoprotein level and coronary disease. *N Engl J Med*. 2009;361:2518–2528. doi: 10.1056/NEJMoa0902604
47. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. Twelve years of SAMtools and BCFtools. *GigaScience*. 2021;10:giab008. doi: 10.1093/gigascience/giab008
48. Pedersen BS, Quinlan AR. Who's who? Detecting and resolving sample anomalies in human DNA sequencing studies with peddy. *Am J Hum Genet*. 2017;100:406–413. doi: 10.1016/j.ajhg.2017.01.017
49. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38:e164. doi: 10.1093/nar/gkq603
50. Frank E, Witten IH. Generating accurate rule sets without global optimization. Paper presented at: Fifteenth International Conference on Machine Learning; July 24–27, 1998; Madison, WI, USA.
51. Cohen WW. Fast effective rule induction. Paper presented at: Twelth International Conference on Machine Learning; July 9–12, 1995; Tahoe City, CA, USA.
52. Ruck DW, Rogers SK, Kabrisky M, Oxley ME, Suter BW. The multilayer perceptron as an approximation to a Bayes optimal discriminant function. *IEEE Trans Neural Netw*. 1990;1:296–298. doi: 10.1109/72.80266
53. John GH, Langley P. Estimating continuous distributions in Bayesian classifiers. Paper presented at: Eleventh Conference on Uncertainty in Artificial Intelligence; August 18–20, 1995; Montreal, Quebec, CA.
54. Cessie SL, Houwelingen JCV. Ridge estimators in logistic regression. *Appl Stat*. 1992;41:191. doi: 10.2307/2347628
55. Aha DW, Kibler D, Albert MK. Instance-based learning algorithms. *Mach Learn*. 1991;6:37–66. doi: 10.1007/BF00153759
56. Breiman L. Random forests. *Mach Learn*. 2001;45:5–32. doi: 10.1023/A:1010933404324
57. Quinlan JR. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers; 1992. doi: 10.11517/jjsai.10.3_475
58. Freund Y, Schapire RE. Experiments with a new boosting algorithm. Paper presented at: Thirteenth International Conference on Machine Learning; July 3–6, 1996; Bari, Italy.
59. Bult CJ, Blake JA, Smith CL, Kadin JA, Richardson JE, Anagnostopoulos A, Asabor R, Baldarelli RM, Beal JS, Bello SM, et al. Mouse genome database (MGD) 2019. *Nucleic Acids Res*. 2019;47:D801–D806. doi: 10.1093/nar/gky1056
60. Lisewski AM, Lichtarge O. Untangling complex networks: risk minimization in financial markets through accessible spin glass ground states. *Phys Stat Mech Appl*. 2010;389:3250–3253. doi: 10.1016/j.physa.2010.04.005
61. Lisewski AM, Quiros JP, Ng CL, Adikesavan AK, Miura K, Putluri N, Eastman RT, Scanfeld D, Regenbogen SJ, Altenhofen L, et al. Supergenomic network compression and the discovery of EXP1 as a glutathione transferase inhibited by artesunate. *Cell*. 2014;158:916–928. doi: 10.1016/j.cell.2014.07.011
62. Pham M, Lichtarge O. Graph-based information diffusion method for prioritizing functionally related genes in protein-protein interaction networks. In: Altman RB, Keith Dunker A, Hunter L, Ritchie MD, Murray T, Klein TE, eds. *Biocomputing 2020*. Kohala Coast, Hawaii, USA: World Scientific; 2019:439–450. doi: 10.1142/9789811215636_0039
63. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, et al. The genotype-tissue expression (GTEx) project. *Nat Genet*. 2013;45:580–585. doi: 10.1038/ng.2653
64. Franzen O, Ermel R, Cohain A, Akers NK, Di Narzo A, Talukdar HA, Foroughi-Asl H, Giambartolomei C, Fullard JF, Sukhavasi K, et al. Cardiometabolic risk loci share downstream cis- and trans-gene regulation across tissues and diseases. *Science*. 2016;353:827–830. doi: 10.1126/science.aad6970
65. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta*. 1975;405:442–451. doi: 10.1016/0005-2795(75)90109-9
66. Watanabe K, Taskesen E, van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun*. 2017;8:1826. doi: 10.1038/s41467-017-01261-5
67. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E, et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. 2019;47:D1005–D1012. doi: 10.1093/nar/gky1120
68. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*. 2018;46:D1062–D1067. doi: 10.1093/nar/gkx1153

69. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, Vilo J. G:profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res*. 2019;47:W191–W198. doi: 10.1093/nar/gkz369
70. van der Harst P, Verweij N. Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circ Res*. 2018;122:433–443. doi: 10.1161/CIRCRESAHA.117.312086
71. Barefield DY, Puckelwartz MJ, Kim EY, Wilsbacher LD, Vo AH, Waters EA, Earley JU, Hadhazy M, Dellefave-Castillo L, Pesce LL, et al. Experimental modeling supports a role for MyBP-HL as a novel myofilament component in arrhythmia and dilated cardiomyopathy. *Circulation*. 2017;136:1477–1491. doi: 10.1161/CIRCULATIONAHA.117.028585
72. Omura T, Yoshiyama M, Ishikura F, Kobayashi H, Takeuchi K, Beppu S, Yoshikawa J. Myocardial ischemia activates the JAK–STAT pathway through angiotensin II signaling in in vivo myocardium of rats. *J Mol Cell Cardiol*. 2001;33:307–316. doi: 10.1006/jmcc.2000.1303
73. Esmon CT. The protein C pathway. *Chest*. 2003;124:26S–32S. doi: 10.1378/chest.124.3_suppl.26s
74. Howson JMM, Zhao W, Barnes DR, Ho W-K, Young R, Paul DS, Waite LL, Freitag DF, Fauman EB, Salfati EL, et al. Fifteen new risk loci for coronary artery disease highlight arterial-wall-specific mechanisms. *Nat Genet*. 2017;49:1113–1119. doi: 10.1038/ng.3874
75. Germain M, Chasman DI, de Haan H, Tang W, Lindström S, Weng L-C, de Andrade M, de Visser MCH, Wiggins KL, Suchon P, et al. Meta-analysis of 65,734 individuals identifies TSPAN15 and SLC44A2 as two susceptibility loci for venous thromboembolism. *Am J Hum Genet*. 2015;96:532–542. doi: 10.1016/j.ajhg.2015.01.019
76. Hariharan P, Dupuis J. Mapping gene and gene pathways associated with coronary artery disease: a CARDIoGRAM exome and multi-ancestry UK biobank analysis. *Sci Rep*. 2021;11:16461. doi: 10.1038/s41598-021-95637-9
77. Lappalainen T, Montgomery SB, Nica AC, Dermitzakis ET. Epistatic selection between coding and regulatory variation in human evolution and disease. *Am J Hum Genet*. 2011;89:459–463. doi: 10.1016/j.ajhg.2011.08.004
78. Castel SE, Cervera A, Mohammadi P, Aguet F, Reverter F, Wolman A, Guigo R, Iossifov I, Vasileva A, Lappalainen T. Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. *Nat Genet*. 2018;50:1327–1334. doi: 10.1038/s41588-018-0192-y
79. Pennacchio LA, Olivier M, Hubacek JA, Cohen JC, Cox DR, Fruchart J-C, Krauss RM, Rubin EM. An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science*. 2001;294:169–173. doi: 10.1126/science.1064852
80. van der Vliet HN, Sammels MG, Leegwater ACJ, Levels JHM, Reitsma PH, Boers W, Chamuleau RAFM. Apolipoprotein A-V: a novel apolipoprotein associated with an early phase of liver regeneration. *J Biol Chem*. 2001;276:44512–44520. doi: 10.1074/jbc.M106888200
81. Defesche JC. Low-density lipoprotein receptor--its structure, function, and mutations. *Semin Vasc Med*. 2004;4:5–11. doi: 10.1055/s-2004-822993
82. Schlange T, Andrée B, Arnold H-H, Brand T. BMP2 is required for early heart development during a distinct time period. *Mech Dev*. 2000;91:259–270. doi: 10.1016/s0925-4773(99)00311-1
83. Ma L, Lu M-F, Schwartz RJ, Martin JF. Bmp2 is essential for cardiac cushion epithelial-mesenchymal transition and myocardial patterning. *Development*. 2005;132:5601–5611. doi: 10.1242/dev.02156
84. Rivera-Feliciano J, Tabin CJ. Bmp2 instructs cardiac progenitors to form the heart-valve-inducing field. *Dev Biol*. 2006;295:580–588. doi: 10.1016/j.ydbio.2006.03.043
85. Fritsche LG, Igl W, Bailey JNC, Grassmann F, Sengupta S, Bragg-Gresham JL, Burdon KP, Hebbring SJ, Wen C, Gorski M, et al. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat Genet*. 2016;48:134–143. doi: 10.1038/ng.3448
86. Kunkle BW, Grenier-Boley B, Sims R, Bis JC, Damotte V, Naj AC, Boland A, Vronskaya M, van der Lee SJ, Amlie-Wolf A, et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat Genet*. 2019;51:414–430. doi: 10.1038/s41588-019-0358-2
87. Joehanes R, Zhang X, Huan T, Yao C, Ying S, Nguyen QT, Demirkale CY, Feolo ML, Sharopova NR, Sturcke A, et al. Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies. *Genome Biol*. 2017;18:16. doi: 10.1186/s13059-016-1142-6
88. Nielsen MJ, Petersen SV, Jacobsen C, Oxvig C, Rees D, Møller HJ, Moestrup SK. Haptoglobin-related protein is a high-affinity hemoglobin-binding plasma protein. *Blood*. 2006;108:2846–2849. doi: 10.1182/blood-2006-05-022327
89. Thomas SA, Matsumoto AM, Palmiter RD. Noradrenaline is essential for mouse fetal development. *Nature*. 1995;374:643–646. doi: 10.1038/374643a0
90. Koyama S, Ito K, Terao C, Akiyama M, Horikoshi M, Momozawa Y, Matsunaga H, Ieki H, Ozaki K, Onouchi Y, et al. Population-specific and trans-ancestry genome-wide analyses identify distinct and shared genetic risk loci for coronary artery disease. *Nat Genet*. 2020;52:1169–1177. doi: 10.1038/s41588-020-0705-3
91. Kurki MI, Karjalainen J, Palta P, Sipilä TP, Kristiansson K, Donner KM, Reeve MP, Laivuori H, Aavikko M, Kaunisto MA, et al. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature*. 2023:613:508–518. doi: 10.1038/s41586-022-05473-8
92. Kim HS, Li A, Ahn S, Song H, Zhang W. Inositol polyphosphate-5-phosphatase F (INPP5F) inhibits STAT3 activity and suppresses gliomas tumorigenicity. *Sci Rep*. 2014;4:7330. doi: 10.1038/srep07330
93. Harhous Z, Booz GW, Ovize M, Bidaux G, Kurdi M. An update on the multifaceted roles of STAT3 in the heart. *Front Cardiovasc Med*. 2019;6:150. doi: 10.3389/fcvm.2019.00150
94. Goldstein JL, Sobhani MK, Faust JR, Brown MS. Heterozygous familial hypercholesterolemia: failure of normal allele to compensate for mutant allele at a regulated genetic locus. *Cell*. 1976;9:195–203. doi: 10.1016/0092-8674(76)90110-0
95. Mendoza-Barberá E, Julve J, Nilsson SK, Lookene A, Martín-Campos JM, Roig R, Lechuga-Sancho AM, Sloan JH, Fuentes-Prior P, Blanco-Vaca F. Structural and functional analysis of APOA5 mutations identified in patients with severe hypertriglyceridemia. *J Lipid Res*. 2013;54:649–661. doi: 10.1194/jlr.M031195
96. Benlian P, De Gennes JL, Foubert L, Zhang H, Gagné SE, Hayden M. Premature atherosclerosis in patients with familial chylomicronemia caused by mutations in the lipoprotein lipase gene. *N Engl J Med*. 1996;335:848–854. doi: 10.1056/NEJM199609193351203
97. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, Ganna A, Chen J, Buchkovich ML, Mora S, et al. Discovery and refinement of loci associated with lipid levels. *Nat Genet*. 2013;45:1274–1283. doi: 10.1038/ng.2797
98. Barton AR, Sherman MA, Mukamel RE, Loh P-R. Whole-exome imputation within UK biobank powers rare coding variant association and fine-mapping analyses. *Nat Genet*. 2021;53:1260–1269. doi: 10.1038/s41588-021-00892-1
99. Smith AB, Esko JD, Hajduk SL. Killing of trypanosomes by the human haptoglobin-related protein. *Science*. 1995;268:284–286. doi: 10.1126/science.7716520
100. Barefield DY, Yamakawa S, Tahtah I, Sell JJ, Broman M, Laforest B, Harris S, Alvarez-Arce A, Araujo KN, Puckelwartz MJ, et al. Partial and complete loss of myosin binding protein H-like cause cardiac conduction defects. *J Mol Cell Cardiol*. 2022;169:28–40. doi: 10.1016/j.yjmcc.2022.04.012
101. Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs KV, Li X, Li H, Kuperwasser N, Ruda VM, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*. 2010;466:714–719. doi: 10.1038/nature09266
102. Song L, Leung C, Schindler C. Lymphocytes are important in early atherosclerosis. *J Clin Invest*. 2001;108:251–259. doi: 10.1172/JCI11380
103. Wolf D, Ley K. Immunity and inflammation in atherosclerosis. *Circ Res*. 2019;124:315–327. doi: 10.1161/CIRCRESAHA.118.313591
104. Liu J, Wang H, Li J. Inflammation and inflammatory cells in myocardial infarction and reperfusion injury: a double-edged sword. *Clin Med Insights Cardiol*. 2016;10:79–84. doi: 10.4137/CMC.S33164
105. Hsu F, Hu F, Mao Y. Spatiotemporal control of phosphatidylinositol 4-phosphate by Sac2 regulates endocytic recycling. *J Cell Biol*. 2015;209:97–110. doi: 10.1083/jcb.201408027
106. Nakatsu F, Messa M, Nández R, Czapla H, Zou Y, Strittmatter SM, De Camilli P. Sac2/INPP5F is an inositol 4-phosphatase that functions in the endocytic pathway. *J Cell Biol*. 2015;209:85–95. doi: 10.1083/jcb.201409064
107. Zhu W, Trivedi CM, Zhou D, Yuan L, Lu MM, Epstein JA. Inpp5f is a polyphosphoinositide phosphatase that regulates cardiac hypertrophic

responsiveness. *Circ Res*. 2009;105:1240–1247. doi: 10.1161/CIRCRESAHA.109.208785

108. Yue H, Li W, Desnoyer R, Karnik SS. Role of nuclear unphosphorylated STAT3 in angiotensin II type 1 receptor-induced cardiac hypertrophy. *Cardiovasc Res*. 2010;85:90–99. doi: 10.1093/cvr/cvp285
109. Wang M-H, Zhou Y-Q, Chen Y-Q. Macrophage-stimulating protein and RON receptor tyrosine kinase: potential regulators of macrophage inflammatory activities. *Scand J Immunol*. 2002;56:545–553. doi: 10.1046/j.1365-3083.2002.01177.x
110. Huang L, Fang X, Shi D, Yao S, Wu W, Fang Q, Yao H. MSP-RON pathway: potential regulator of inflammation and innate immunity. *Front Immunol*. 2020;11:1–8. doi: 10.3389/fimmu.2020.569082
111. Waltz SE, Eaton L, Toney-Earley K, Hess KA, Peace BE, Ihlendorf JR, Wang M-H, Kaestner KH, Degen SJF. Ron-mediated cytoplasmic signaling is dispensable for viability but is required to limit inflammatory responses. *J Clin Invest*. 2001;108:567–576. doi: 10.1172/JCI11881
112. Lin H-B, Naito K, Oh Y, Farber G, Kanaan G, Valaperti A, Dawood F, Zhang L, Li GH, Smyth D, et al. Innate immune Nod1/RIP2 signaling is essential for cardiac hypertrophy but requires mitochondrial antiviral signaling protein for signal transductions and energy balance. *Circulation*. 2020;142:2240–2258. doi: 10.1161/CIRCULATIONAHA.119.041213
113. Guella I, Duga S, Ardissino D, Merlini P, Peyvandi F, Mannucci P, Asselta R. Common variants in the haemostatic gene pathway contribute to risk of early-onset myocardial infarction in the Italian population. *Thromb Haemost*. 2011;106:855–864. doi: 10.1160/TH11-04-0247
114. Tang W, Basu S, Kong X, Pankow JS, Aleksic N, Tan A, Cushman M, Boerwinkle E, Folsom AR. Genome-wide association study identifies novel loci for plasma levels of protein C: the ARIC study. *Blood*. 2010;116:5032–5036. doi: 10.1182/blood-2010-05-283739
115. Smith NL, Chen M-H, Dehghan A, Strachan DP, Basu S, Soranzo N, Hayward C, Rudan I, Sabater-Lleal M, Bis JC, et al. Novel associations of multiple genetic loci with plasma levels of factor VII, factor VIII, and von Willebrand factor. *Circulation*. 2010;121:1382–1392. doi: 10.1161/CIRCULATIONAHA.109.869156
116. Dennis J, Johnson CY, Adediran AS, de Andrade M, Heit JA, Morange P-E, Trégouët D-A, Gagnon F. The endothelial protein C receptor (PROCR) Ser219Gly variant and risk of common thrombotic disorders: a HuGE review and meta-analysis of evidence from observational studies. *Blood*. 2012;119:2392–2400. doi: 10.1182/blood-2011-10-383448
117. Ong YS, Tang BL, Loo LS, Hong W. p125A exists as part of the mammalian Sec13/Sec31 COPII subcomplex to facilitate ER-Golgi transport. *J Cell Biol*. 2010;190:331–345. doi: 10.1083/jcb.201003005
118. An H, Ordureau A, Paulo JA, Shoemaker CJ, Denic V, Harper JW. TEX264 is an endoplasmic reticulum-resident ATG8-interacting protein critical for ER remodeling during nutrient stress. *Mol Cell*. 2019;74:891–908.e10. doi: 10.1016/j.molcel.2019.03.034
119. Chino H, Hatta T, Natsume T, Mizushima N. Intrinsically disordered protein TEX264 mediates ER-phagy. *Mol Cell*. 2019;74:909–921.e6. doi: 10.1016/j.molcel.2019.03.033
120. Vujkovic M, Keaton JM, Lynch JA, Miller DR, Zhou J, Tcheandjieu C, Huffman JE, Assimes TL, Lorenz K, Zhu X, et al. Discovery of 318 new risk loci for type 2 diabetes and related vascular outcomes among 1.4 million participants in a multi-ancestry meta-analysis. *Nat Genet*. 2020;52:680–691. doi: 10.1038/s41588-020-0637-y
121. Richardson TG, Sanderson E, Palmer TM, Ala-Korpela M, Ference BA, Davey Smith G, Holmes MV. Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: a multivariable Mendelian randomisation analysis. *PLoS Med*. 2020;17:e1003062. doi: 10.1371/journal.pmed.1003062
122. Chang C-C, Wu C-H, Liu L-K, Chou R-H, Kuo C-S, Huang P-H, Chen L-K, Lin S-J. Association between serum uric acid and cardiovascular risk in nonhypertensive and nondiabetic individuals: the Taiwan I-Lan Longitudinal Aging Study. *Sci Rep*. 2018;8:5234. doi: 10.1038/s41598-018-22997-0
123. Borghi C, Piani F. Uric acid and risk of cardiovascular disease: a question of start and finish. *Hypertension*. 2021;78:1219–1221. doi: 10.1161/HYPERTENSIONAHA.121.17631
124. Creider JC, Hegele RA, Joy TR. Niacin: another look at an underutilized lipid-lowering medication. *Nat Rev Endocrinol*. 2012;8:517–528. doi: 10.1038/nrendo.2012.22
125. Hosseini L, Vafaee MS, Badalzadeh R. Melatonin and nicotinamide mononucleotide attenuate myocardial ischemia/reperfusion injury via modulation of mitochondrial function and hemodynamic parameters in aged rats. *J Cardiovasc Pharmacol Ther*. 2020;25:240–250. doi: 10.1177/1074248419882002