Journal of
Personalized
Medicine

MDPI

Article

Genome Wide Epistasis Study of On-Statin Cardiovascular Events with Iterative Feature Reduction and Selection

Solomon M. Adams *, Habiba Feroze †, Tara Nguyen †, Seenae Eum, Cyrille Cornelio and Arthur F. Harralson

Department of Pharmacogenomics, Shenandoah University School of Pharmacy, Fairfax, VA 22031, USA; hferoze17@su.edu (H.F.); tnguyen174@su.edu (T.N.); seum@su.edu (S.E.); ccorneli1@su.edu (C.C.); aharrals@su.edu (A.F.H.)

* Correspondence: sadams07@su.edu; Tel.: +1-540-542-6237

† These authors contributed equally to this work.

Received: 10 September 2020; Accepted: 4 November 2020; Published: 7 November 2020

Abstract: Predicting risk for major adverse cardiovascular events (MACE) is an evidence-based practice that incorporates lifestyle, history, and other risk factors. Statins reduce risk for MACE by decreasing lipids, but it is difficult to stratify risk following initiation of a statin. Genetic risk determinants for on-statin MACE are low-effect size and impossible to generalize. Our objective was to determine high-level epistatic risk factors for on-statin MACE with GWAS-scale data. Controlled-access data for 5890 subjects taking a statin collected from Vanderbilt University Medical Center's BioVU were obtained from dbGaP. We used Random Forest Iterative Feature Reduction and Selection (RF-IFRS) to select highly informative genetic and environmental features from a GWAS-scale dataset of patients taking statin medications. Variant-pairs were distilled into overlapping networks and assembled into individual decision trees to provide an interpretable set of variants and associated risk. 1718 cases who suffered MACE and 4172 controls were obtained from dbGaP. Pathway analysis showed that variants in genes related to vasculogenesis (FDR = 0.024), angiogenesis (FDR = 0.019), and carotid artery disease (FDR = 0.034) were related to risk for on-statin MACE. We identified six gene-variant networks that predicted odds of on-statin MACE. The most elevated risk was found in a small subset of patients carrying variants in *COL4A2*, *TMEM178B*, *SZT2*, and *TBXAS1* (OR = 4.53, $p < 0.001$). The RF-IFRS method is a viable method for interpreting complex "black-box" findings from machine-learning. In this study, it identified epistatic networks that could be applied to risk estimation for on-statin MACE. Further study will seek to replicate these findings in other populations.

Keywords: pharmacogenomics; epistasis; random forest; statin; cardiovascular disease

1. Introduction

Predicting risk for Cardiovascular Disease (CVD) is a mainstay of primary care and cardiology. Patients who develop CVD are at risk for major adverse cardiovascular events (MACE), such as myocardial infarction, stroke, or unstable angina. Risk assessments for CVD include clinical biomarkers, family history, lifestyle, co-morbidities and biometrics. Routine risk assessments for CVD risk guide major therapeutic and lifestyle decisions.

Hyperlipidemia is a risk factor for CVD and MACE, and the American College of Cardiology (ACC) guidelines on the management of blood cholesterol recommend statins as the cornerstone pharmacotherapy [1]. CVD risk reduction from statins might be population specific and shows diversity among different patient groups. Ramos and colleagues found that the incidence of MACE was 19.7 (statin-users) and 24.7 (statin non-users) events per 1000 person-years in patients with asymptomatic peripheral artery disease [2]. Another study concluded that statin therapy had no major benefit on stroke in women [3]. Overall, however, statins reduce the risk for MACE proportional to the magnitude of cholesterol lowering in all ages [4] .

The clinical pharmacogenetics implementation consortium guidelines support the use of pharmacogenomics (PGx) assessment for prevention of myopathy with simvastatin based on patients' *SLCO1B1* genotype [5]. Additionally, statin biochemical response (e.g., PK, Lipid Lowering Efficacy) is associated with numerous genomic variations. Ruiz-Iruela and colleagues found that decreased lipid lowering of rosuvastatin, atorvastatin, and simvastatin is predicted by *ABCA1* rs2230806 and *CYP2D6*. They also found that *CETP* variants rs708272 and rs5882 were associated with decreased and increased LDL lowering with rosuvastatin, respectively [6]. These variations, however, have not been found to be associated with higher level outcomes like prevention of CVD-related events. Low-effect size risk variants also provide insight into pathogenesis of CVD. Genetic variations in apolipoprotein C-III (*APOC3*) and angiopoietin-like 4 (*ANGPTL4*) have been associated with risk for coronary artery disease (CAD) [7]. Roguin and colleagues found that the Haptoglobin (*HP*) genotype was a significant independent predictor of MACE in patients with diabetes [8]. The PROSPER study found that *SURF6* rs579459 was associated with CAD, stroke and large artery stroke. It also found that *TWIST1* rs2107595 was associated with an increased risk of MACE such as large artery stroke, CAD, and ischemic stroke [9]. Routine genetic testing for hyperlipidemia and CVD risk is limited to patients with history of familial hypercholesterolemia (FH), predicted by variation in *LDLR*, *APOB*, or *PCSK9* [1]. CVD nevertheless shows strong heritability in patients without FH, suggesting an underlying genetic component [7]. Genome-Wide Association Studies (GWAS) have identified over 50 genetic variants that are associated with risk for CVD and MACE. Clinical translation of these genetic risk factors is challenged by individual variants with small effect sizes and poor understanding of the interplay between multiple genetic variants and risk for MACE. These genetic factors might help explain cases in which patients still experience MACE in spite of adequate phamacologic response to statin therapy and other risk reduction strategies.

PGx is exemplified in variations among drug-metabolizing genes, including phase I (oxidation, reduction, hydrolysis), phase II (conjugation), and phase III (transport). In these cases, functional genetic variations can have catastrophic effects on pharmacokinetics [10]. While some evidence supports PGx for pharmacodynamic markers, PGx outside of pharmacokinetics has been limited by relatively low effect-size of individual variants, and the inability to consistently apply multiple gene effects. This is partially addressed by the growing use of polygenic risk scores (PRS) to pool effects from unrelated variants [11]; however, little has been done to incorporate the effects of epistasis (i.e., gene-gene interactions) to create novel predictors of drug response.

The objective of this research was to stratify the risk of on-statin MACE based on polygenic epistatic predictors. We applied a step-wise, interpretable, machine-learning (ML) driven ensemble method for feature reduction and determination of epistasis to a GWAS-scale dataset. We expect that application of this method will drive novel insight into genetic interactions that drive risk for complex cardiovascular phenotypes and statin PGx.

2. Results

2.1. Demographics

Demographic data are summarized in Table 1. Our analysis incorporated genetic variant data and sex, which were available for all subjects. Random forest models do not tolerate missing values and require either imputation or exclusion to include variables with missing data. Given the focus on epistasis in this analysis, non-genetic variables were only included if they were defined in all cases and controls. Weight, and height were frequently missing in controls, and were therefore not used.

Age was only available as "age of first event", which limited its utility in comparing cases and controls. More than 99% of subjects in this population are reported as white. Sex is also a well-established predictor of risk for MACE and was thus included in the model.

Table 1. Population demographics.

Variable	Control	Case	*p*
Female(%)	38.0%	31.2%	<0.001
White(%)	99.3%	99.4%	0.507
BMI(Mean ± SD)	29.03 ± 7.37	28.57 ± 7.035	0.253
Age First MACE (Median ± IQR)	N/A	65 ± 16	1

2.2. Feature Selection with RF-IFRS

After pruning, there were 637,732 variants and 5890 subjects in the cohort. Of the subjects, there were 1718 cases and 4172 controls. Evaluation of additive statistical association did not identify any variants that met genome-wide significance. The RF with the corrected impurity importance measure identified 6688 variants with a corrected-impurity *p* value less than 0.01. As with statistical association, no variants met genome-wide significance. The 6688 initially selected variants were extracted from the full dataset and analyzed with r2VIM. This identified 49 genetic variants in addition to sex with a minimum permutation importance value of at least one. Results from these analyses are shown in Figure 1.

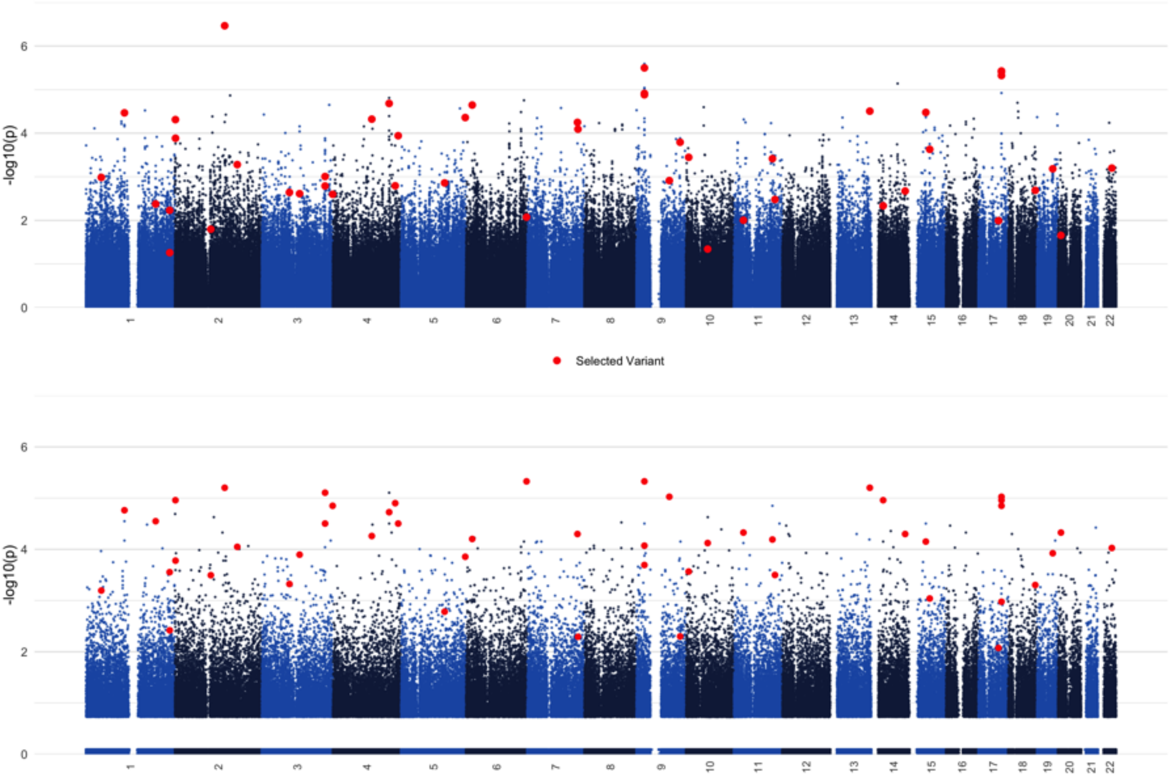

Figure 1. Manhattan plots for statistical GWAS analysis with PLINK (**top**) vs. the initial RF model with ranger (**bottom**). Red dots correspond to variants that were selected with r2VIM, and show that a purely statistical approach fails to identify variants that are likely relevant to the outcome due to interactions.

2.3. Epistasis Screening

Paired selection frequency results identify variant-pairs that co-occur in decision trees more often, as often, or less often than predicted based on individual variant selection. Variants that are selected together more often than expected suggests a greater phenotype prediction from both variants together, and selection less often than expected suggests that co-occurrence comes at a cost to phenotype prediction (i.e., variants are correlated and/or in linkage disequilibrium).

Figure 2 shows the distribution of expected tree co-occurrence for each variant pair. Using an alternative hypothesis of "greater than expected" in a binomial test allows sensitive selection of variant pairs that are chosen more often than predicted (red). We found evidence of epistasis in 16 variant-variant pairs Table 2. Additionally, five variants showed significant interaction with sex.

Condensing variant-pairs based on overlap resulted in six variant networks Figure 3. We found networks that involved intergenic variants, for which the functional consequence is not clear. This is evident in network 1, where sex precedes four intergenic variants, most of which are more than 100 kb away from the nearest gene. Gene-variant networks show diversity in odds for experiencing MACE, with individual node odds ratios reflecting the contribution of multiple variant effects through additive and non-additive relationships.

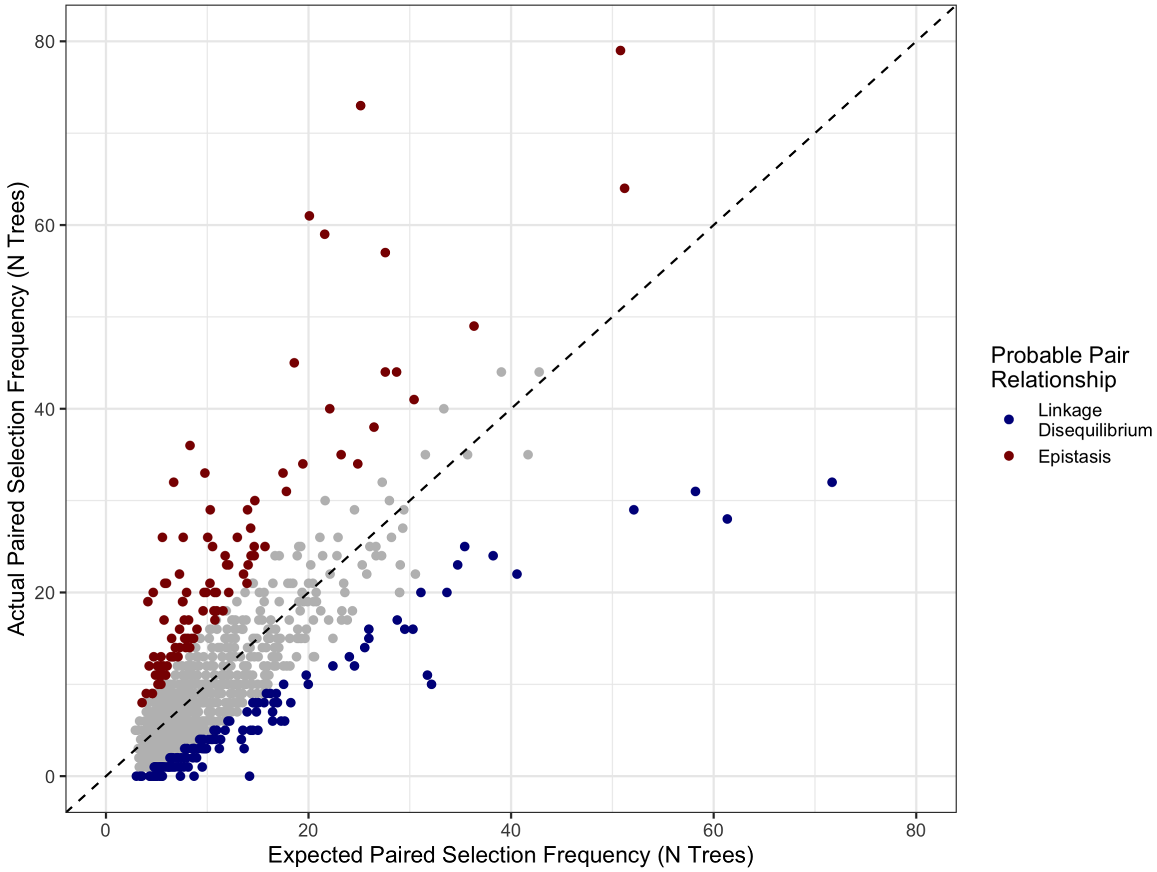

Figure 2. Paired selection frequency based on the combined independent variant probabilities (X axis) vs. the actual frequency of variants being selected together in a decision tree. Variants that are selected together at a lower-than-expected frequency are expected to be correlated with respect to the outcome, suggesting that they are in linkage disequilibrium (blue). Variants selected together more often than expected (red) are predicted to exhibit epistasis with respect to the phenotype.

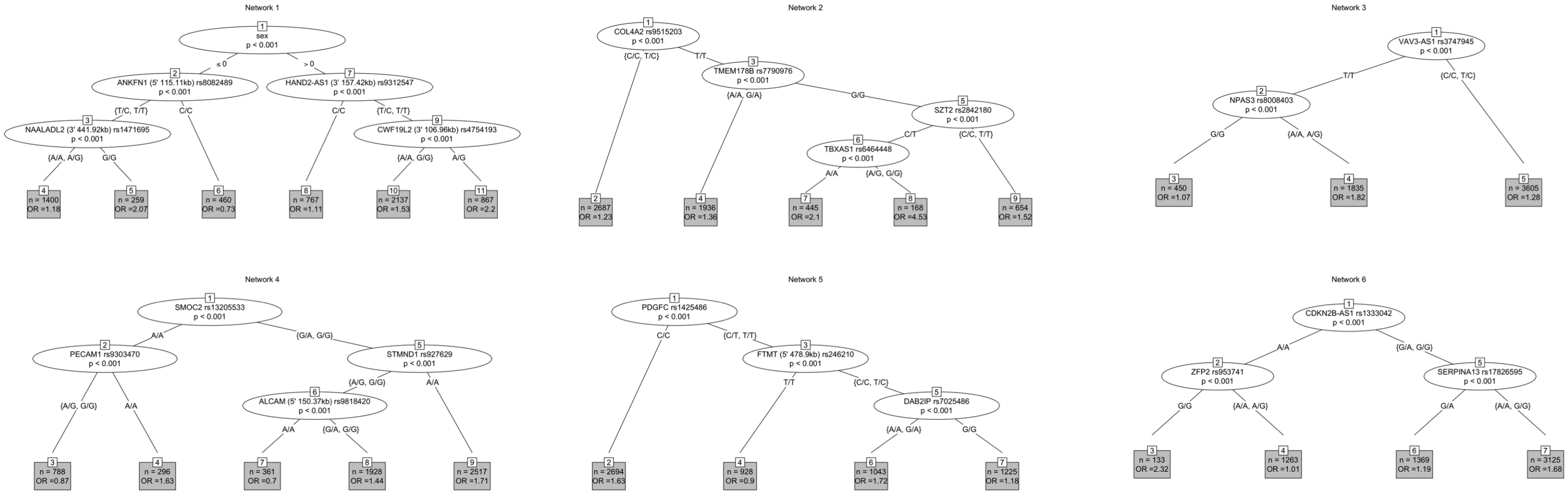

Figure 3. Decision trees incorporating overlapping epistasis-variant-pairs show six unique networks of genes and variants. Odds ratios in terminal nodes represent subject odds of on-statin MACE in someone carrying the collection of alleles shown in the network relative to those who did not carry those variants. This shows a practical interpretation of epistasis findings that might be more practical to incorporate into clinical practice, though validation and replication in independent populations will be necessary to drive clinical translation.

Table 2. Significant Ensemble and Regression Variant Pairs.

Variant 1	Variant 2	$FDR_{ensemble}$	$FDR_{interaction}$
sex	CDCA7 (3′ 242.48 kb) rs6731912	0.001	0.029
sex	NAALADL2 (3′ 441.92 kb) rs1471695	<0.001	0.082
sex	HAND2-AS1 (3′ 157.42 kb) rs9312547	<0.001	0.007
sex	NNMT (5′ 4.01 kb) rs2244175	0.021	0.016
sex	ANKFN1 (5′ 115.11 kb) rs8082489	<0.001	0.007
SZT2 rs2842180	COL4A2 rs9515203	<0.001	0.004
VAV3-AS1 rs3747945	NPAS3 rs8008403	0.001	0.011
KCNT2 (3′ 1239.56 kb) rs6693848	PECAM1 rs2812	<0.001	0.004
KCNT2 (3′ 1239.56 kb) rs6693848	PECAM1 rs9303470	<0.001	0.004
KCNT2 (3′ 1239.56 kb) rs6693848	PECAM1 (5′ 1.22 kb) rs6504218	0.032	0.004
ALCAM (5′ 150.37 kb) rs9818420	STMND1 rs927629	<0.001	0.001
NAALADL2 (3′ 441.92 kb) rs1471695	RFX7 (5′ 10.73 kb) rs2713935	<0.001	0.005
PDGFC rs1425486	FTMT (5′ 478.9 kb) rs246210	<0.001	0.001
FTMT (5′ 478.9 kb) rs246210	DAB2IP rs7025486	<0.001	0.001
ZFP2 rs953741	CDKN2B-AS1 rs1333042	0.011	0.004
STMND1 rs927629	SMOC2 rs13205533	<0.001	0.004
SMOC2 rs13205533	PECAM1 rs2812	0.043	0.016
TBXAS1 rs6464448	COL4A2 rs9515203	0.014	0.009
TMEM178B rs7790976	COL4A2 rs9515203	0.043	0.004
CDKN2B-AS1 rs2383207	SERPINA13 rs17826595	0.001	0.016
SFMBT2 rs10453997	CWF19L2 (3′ 106.96 kb) rs4754193	<0.001	0.001
CWF19L2 (3′ 106.96 kb) rs4754193	NNMT (5′ 4.01 kb) rs2244175	0.006	0.011
GATM (3′ 12.69 kb) rs2461700	ZNF404 rs1978723	<0.001	0.005

We found that a small subset of subjects carrying *COL4A2* rs9515203 (T/T), *TMEM178B* rs7790976 (G/G), *SZT2* rs2842180 (C/T), and *TBXAS1* rs6464448 (G Allele) showed the highest increase in MACE risk (Network 2, OR = 4.53, $p < 0.001$). Variant effects analysis showed evidence of gene networks associated with angiogenesis, endothelial cell development and function, carotid artery disease, and development of vasculature (minimum FDR = 0.019) Table 3.

Table 3. Gene network associated disease processes

Diseases or Functions	Genes	FDR
Angiogenesis	ALCAM CDKN2B COL4A2 DAB2IP PDGFC PECAM1 SMOC2 VAV3	0.0188
Carotid artery disease	NNMT VAV3	0.034
Development of vasculature	ALCAM CDKN2B COL4A2 DAB2IP NPAS3 PDGFC PECAM1 SMOC2 VAV3	0.0188
Endothelial cell development	COL4A2 PDGFC PECAM1 SMOC2	0.0291
Formation of blood vessel	CDKN2B COL4A2 PECAM1	0.0242
Formation of endothelial tube	COL4A2 PECAM1	0.0291
Function of endothelial tissue	PECAM1 VAV3	0.0188
Migration of endothelial cells	ALCAM COL4A2 PECAM1 SMOC2 VAV3	0.0188
Quantity of endothelial cells	ALCAM PDGFC	0.023
Vasculogenesis	ALCAM CDKN2B COL4A2 PDGFC PECAM1 SMOC2	0.0242

3. Discussion

3.1. Risk Variants and Interactions for CVD

This study was a genome-wide study for variant-variant interactions (epistasis) associated with on-statin MACE. We found six variant networks that show a diverse range of genetic interactions that predict increased or decreased risk for on-statin MACE. Our findings show that RF-IFRS produces polygenic predictors of risk for on-statin MACE, suggesting that limitations of low effect-sizes can be overcome by studying variant networks to produce a final odds ratio.

3.2. Novelty and Application to Clinical Practice

Machine learning techniques based on RF have been commonplace in the evaluation of gene-gene interactions in genomic data [12]. The novelty of the RF-IFRS method primarily derives from the direct analysis of forest structure to estimate epistasis, with application to clinical data. This method shows similarity to work by Li and colleagues, who used a permutation-based RF method to find networks of gene-gene interactions in simulated and real data [13]. While our approach is similar, RF-IFRS scales to GWAS-sized data and corrects for case-control imbalance and variable allele frequency, which is a challenge in many other RF implementations [14]. These studies should ultimately seek clinical translation, and is reflected in this study of a highly relevant clinical phenotype (on-statin MACE). Nevertheless, it is critical that readers recognize that this work and gene/variant-interaction networks are preliminary and have not been evaluated *in vitro*.

3.3. Angiogenesis, Endothelial Function, and Vasculogenesis in CVD

Angiogenesis refers to the formation of new capillary beds from existing vasculature, whereas vasculogenesis refers to the formation of *de novo* vascular networks (i.e., during embryonic development) [15]. Among others, *ALCAM, CDKN2B, COL4A2, DAB2IP, PECAM1, SMOC2, VAV3,* and *PDGFC* are related to either angiogenesis and/or vasculogenesis. These genes were included in five out of six networks that we identified, suggesting that these processes are relevant to risk for MACE and on-statin MACE.

3.4. RF-IFRS Replicates Existing Gene Associations with CVD and Incorporates Novel Interactions

Network one incorporated interactions with sex and variants in four intergenic regions. These variants flanked the nearest genes (*NAALADL2, HAND2-AS1, NNMT,* and *ANKFN1*) by up to 450 kb. Drawing mechanistic insight from these interactions is not practical or necessarily advisable without further mechanistic analysis. However, this finding suggests that the association of male sex with higher risk for on-statin MACE is connected to diverse genetic components that might connect to chromatin structure, un-annotated regulatory RNA genes (e.g., lncRNA, Micro-RNA, etc).

Network two shows a relationship between *COL4A2* rs9515203, *TMEM178B* rs7790976, *SZT2* rs2842184, and *TBXAS1* rs6464448. *COL4A2*. (Collagen Type IV Alpha 2 Chain) codes for the collagen IV peptide α 2 chain, which is a component of the basement membrane surrounding the endothelium of blood vessels [16]. *COL4A2* rs9515203 has previously shown association with sub-clinical atherosclerosis [17], and coronary artery disease [18,19]. Other variants in *COL4A2* and *COL4A1* show associations with risk for MI, atheroslcerotic plaque stability, and vascular stability [16]. The role of *SZT2* (Seizure threshold 2 homolog) rs2842184 in CVD is not clear, and may not indicate a direct mechanism. A recent proteomic study of plasma protein expression in patients with CVD found decreased plasma levels of SZT2 in patients with CVD. The authors suggested that this might be connected to increased mTORC1 signalling in patients with CVD, but this mechanism has not been tested [20]. *TMEM178B* (Transmembrane Protein 178B) codes for a transmembrane protein that is highly expressed in cardiac tissue, among others. The role of the rs7790976 variant is not clear in this network. *TBXAS1* (Thromboxane A synthase 1) codes for Thromboxane A Synthase 1, which is expressed in several tissues including platelets. Thromboxane is a potent vasoconstrictor that causes vasoconstriction and platelet aggregation. The rs6464448 variant has not been previously associated with a phenotype, and the role of genetic variation connecting *TBXAS1* to CVD outcomes is not clear. However, *TBXAS1* has been recently proposed as a potential drug target for CVD [21].

Network three is comprised of an interaction between *VAV3-AS1* rs3747945 and *NPAS3* rs8008403. *VAV3* (Vav Guanine Nucleotide Exchange Factor 3) is important to the migration of smooth muscle cells, which suggests that it has a role in vascular proliferation [22]. *VAV3-AS1* is an RNA gene coding for anti-sense *VAV3*, which might regulate expression of *VAV3* [23]. *VAV3-AS1* rs3747945 has not been previously associated with cardiovascular disease related outcomes, but further supports the role for vasculogenesis in risk for MACE. *NPAS3* (Neuronal PAS Domain Protein 3) rs8008403 has not been previously associated with cardiovascular disease related outcomes, but another variant in *NPAS3* (rs17460823) was associated with C-reactive protein in patients taking fenofibrate [24]. The mechanistic connection of these variants/genes is difficult to determine, but might be linked to development of gross anatomy of the cardiovascular system, or to remodeling associated with CVD.

Network four connects *SMOC2* rs13205533, *PECAM1* rs9303470, *STMND1* rs927629, and a variant (rs9818420) approximately 150 kb upstream from *ALCAM*. This network appears to be related to vascular homeostasis and proliferation. *SMOC2* (SPARC-related modular calcium-binding protein 2) modulates calcium homeostasis, and might be relevant to blood vessel calcification [25]. *SMOC2* rs13205533 has not been previously associated with cardiovascular disease related outcomes. *PECAM1* (Platelet And Endothelial Cell Adhesion Molecule 1) is important for the maintenance of vascular endothelial integrity, and endothelial cells that express PECAM1 are more resilient to the inflammatory response from vascular barrier damage [26,27]. *PECAM1* rs9303470 has not been previously associated with cardiovascular disease related outcomes, but other variants in *PECAM1* have been found to be associated with CAD [26]. *PECAM1* shares similar function with *ALCAM* (Activated leukocyte cell adhesion molecule), and both seem to play roles in CVD [28]. Higher levels of the ALCAM protein have been associated with poor CV outcomes including CV death in patients presenting with ACS [28]. *STMND1* (Stathmin Domain Containing 1) Variants in *STMND1* have been associated with stroke in African Americans, though rs927629 has not been previously reported with CVD [29].

Network five shows interactions between genes relevent to angiogenesis, including *PDGFC* rs1425486 and *DAB2IP* rs7025486. Variants in *PDGFC* (Platelet Derived Growth Factor C) and other *PDGF* genes have been associated with angiogenesis and CVD [30]. PDGFC likely promotes angiogenesis independently of VEGF, which might support a role in CVD development and/or vascular remodeling [31]. *PDGFC* rs1425486 has not been previously associated with cardiovascular disease related outcomes. *DAB2IP* (DAB2-interacting protein) is expressed widely in the cardiovascular system and it is believed to be an inhibitor of VEGF-2 signalling and thus an inhibitor of angiogenesis [32]. Multiple variants in *DAB2IP* have been associated with CAD,[33] and rs7025486 is associated with abdominal aortic aneurysm [34].

Network six includes interactions between *CDKN2B-AS1* rs1333042, *ZFP2* rs953741, and *SERPINA13* rs17826595. *CDKN2B-AS1* (cyclin-dependent kinase inhibitor 2B antisense RNA 1) is an RNA gene that regulates the expression of *CDKN2B*. CDKN2B is an inhibitor of cellular proliferation, though its direct role in CVD is not clear. Numerous variants in *CDKN2B-AS1*, including rs1333042, have been associated with CHD [35]. *ZFP2* (Zinc Finger Protein) is a regulator protein. Variants in *ZFP2* are associated with MI in African Americans[29], though *ZFP2* rs953741 has not been previously associated with cardiovascular disease related outcomes. *SERPINA13* (Serpin Family A Member 13) is a pseudogene, and it is not clear what its role is in CVD. *SERPINA13* rs17826595 has not been previously associated with cardiovascular disease related outcomes. Other members of the *SERPIN* gene superfamily are related to cardiovascular system development and regulation [36].

3.5. Limitations

The RF-IFRS method is a novel approach to genome-wide epistasis that incorporates statistics and interpretable ML methods. The definition of MACE used in this study is less broad than is commonly used in the CVD literature. Notably, ischemic stroke and CV death are not included in the definitions, which is relevant to the generalizability of these findings to other studies that evaluate MACE as an outcome. This study was carried out in a single cohort of patients without replication, however, the RF procedure performs thousands of random samples from the dataset to determing feature importance. While this is not as robust as independent replication, it might help mitigate the bias associated with genetic association studies carried out in a single cohort. We did not split the cohort into training and testing groups or perform hyperparameter tuning, which are often done when developing a predictive

ML model. However, the objective was not to generate a highly predictive ML model, but rather to use the organic structure of the RF approach to identify important variants and interactions. This is also relevant to statistical power, and given that this study found no significant variants with traditional GWAS we opted to keep the entire cohort together to maximize power. Due to the RF inclination to discover LD organically, we did not perform LD pruning. We also did not perform imputation to limit the computational overhead required for the RF model training. This study does not include causal analysis of individual SNPs, thus we do not suggest that the reported variants are necessarily causal. Finally, we did not have access to more extensive clinical data. Further analysis and replication ought to evaluate if findings correspond to degree of lipid control.

4. Materials and Methods

4.1. Clinical Dataset

The data/analyses presented in the current publication are based on the use of controlled-access study data downloaded with permission from the dbGaP web site, under phs000963.v1.p1 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000963.v1.p1). This dataset was assembled through Vanderbilt University Medical Center's BioVU repository and clinical data was extracted from the electronic medical record. All subjects gave their informed consent for inclusion before they participated in the study. The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Vanderbilt Institutional Review Board. These data correspond to 5890 subjects of European descent taking HMG-Coa Reductase Inhibitors (statins) who were genotyped with the Illumina HumanOmniExpressExome 8v1-2_A array by the RIKEN Integrative Medical Sciences Center (IMS) and supported by the Pharmacogenomics Research Network (PGRN)-RIKEN IMS Global Alliance. Inclusion and exclusion criteria for cases and controls is described in Table 4. The primary outcome is on-statin MACE, defined as any revascularization event (e.g., stent placement, bypass) and/or acute myocardial infarction. The case group contains 1718 subjects, and the control group contains 4172 subjects. Case and control status was determined with Vanderbilt's BioVU DNA databank and associated Synthetic Derivative database of clinical information, and software tools developed to identify drugs and clinical events using Electronic Health Record-derived structured and unstructured ("free text") data.

Table 4. Inclusion and Exclusion Criteria.

MACE on statin, defined as either AMI or revascularization on statin
AMI on statins: Case definition (all three conditions required): - At least two ICD9 code for AMI or other acute and subacute forms of ischemic heart disease within a five-day window - Confirmed lab within the same time window - Statin prescribed prior to the AMI event in medical records at least 180 days *Revascularization while on statin: Case definition (both conditions required):* - At least one revascularization CPT code - Statin prescribed prior to the revascularization event in medical records at least 180 days
Case Exclusion: - No diagnosis code for AMI, other acute and subacute forms of ischemic heart disease, or historical AMI assigned previously - No revascularization CPT codes assigned previously - No MACE (Major Adverse Cardiovascular Events) found in previous problem list by NLP
Control definition: - Statin prescribed - No diagnosis code for AMI, other acute and subacute forms of ischemic heart, or historical AMI assigned previously - No revascularization CPT codes assigned previously - No MACE found in previous problem list by NLP - Controls match cases by age, gender, statin type (e.g., simvastatin), and statin exposure

4.2. Data Pre-Processing

Data obtained from dbGaP were in Plink format. The XY pseudo-autosomal region was recoded, and then the resulting file was converted to a multi-sample VCF file. The VCF file chromosome and positions were recoded based on the Illumina variant IDs from the HumanOmniExpressExome manifest file Infinium OmniExpressExome-8 v1.6 for GRCh38. Positions with ambiguous chromosome or positions were filtered from the resulting VCF file. Finally, filters were applied so that variants included in the final analysis were autosomal variants with a minor allele frequency of at least one percent. A PLINK format phenotype file was created from the original phenotype file from dbGaP. This created the necessary ID columns and selects the phenotype column corresponding to MACE. The resulting VCF file was converted to the transposed PLINK (tped) format with PLINK and carried forward for additional analyses.

4.3. Random Forest Iterative Feature Reduction and Selection (RF-IFRS)

Code corresponding to methods is available at https://github.com/sadams-lab/manuscript_onstatin-mace-GWES. PLINK format files were read into an R environment with the GenABEL package [37]. To account for the sensitivity of random forest (RF) models to group imbalance, we weighted cases and controls so that the probability of selecting either from a bootstrapped population was equivalent. To provide a reference comparison, we performed genome-wide association (GWA) analysis on the data with Plink version 1.9 using an additive model with no covariates [38]. We used a two-step process for feature selection that sought to overcome computational limits of analyzing highly dimensional GWAS data. The first stage of feature reduction was performed using the Ranger package for R, in which the forest was grown with a mtry fraction of 1/3, and 1000 trees [39] Considering that this method incorporates the full breadth of data, we used the corrected impurity score implemented by Nembrini and colleagues, which overcomes the sensitivity of GINI importance to allele frequency while allowing a practical computing time compared to the more robust permutation score [14]. This method is computationally fast, but relatively non-specific and produces false-positives similar to that of a traditional GWAS.

Features with p values of < 0.01 were selected for secondary feature selection with r2VIM, which incorporates multiple RF models to build a consensus permutation importance [40]. It was re-implemented by Degenhardt and colleagues to support the ranger package, which allows for parallel tree building and much faster execution in the Pomona package [41]. For our implementation, we cloned the Pomona repository and modified it so that it would accept input from a GenABEL object. The resulting custom r2VIM implementation was run with 11 sequentially grown RF models with 10,000 trees per forest using, an mtry fraction of 1/3, and nodes were limited to a maximum of 10% of the total population to limit tree depth. Features from the first forests with a minimum permutation importance of at least one in each forest were selected for estimation of association and interaction. The final (11th) forest was saved for the ensemble-method for epistasis selection.

4.4. Testing for Epistasis

The ensemble method for epistasis estimation was implemented based on the work by Schmalohr and colleagues [42]. We implemented methods for testing paired selection frequency (i.e., the probability that a variable will be included in the same decision tree) and selection asymmetry (i.e., the probability of a variant favoring a particular node when following another variant) [42]. These methods provides the means to detect AND and XOR epistasis. To create a final estimate for the presence of an interaction, p values from each method were combined using the Fisher method [43].

Variant-pair p values were adjusted with the Benjamini-Hochberg FDR method, and pairs with FDR of less than 0.05 were retained for further analysis [44]. Selected variant pairs were converted to dummy variables, and all pair-wise genotype permutations were compared with logistic regression. The minimum pairwise interaction p value was retained for each variant pair. Interaction p values were adjusted with the Benjamini-Hochberg FDR method.

4.5. Poly-Epistatic Risk and Pathway Analysis

To extend beyond pair-wise interactions, pairwise interacting variants were condensed based on overlap. For example, $A|B$ and $A|C$ -> A | B | C. Decision trees were built from the resulting variant interaction networks to visualize relationships and odds ratios based on multiple variants. Decision trees were built with the ctree function in the Party package for R [45] Odds ratios for terminal nodes were normalized to the overall odds of being a case.

To incorporate basic mechanistic insight, data were analyzed through the use of Ingenuity® Variant Analysis^{TM} version 1.18.06(https://www.qiagenbioinformatics.com/products/ingenuity-variant-analysis) from QIAGEN, Inc. (Hilden, Germany). Top diseases and bio-functions relevant to MACE and CVD were reported with correlation to identified decision trees, then filtered for at least two genes involved and a FDR corrected p value of less than 0.05.

5. Conclusions

A RF driven method for feature reduction and selection applied to a GWAS-scale dataset identified six epistasis-networks that may provide insight into the risk for on-statin MACE. This method also provides interpretable results, which may produce a more physiologically relevant assessment of odds and risk for an outcome than PRS. We found that variants related to angiogenesis and vasculogenesis are associated with odds of on-statin MACE. These findings present a unique opportunity for the incorporation of multiple low-effect size variants in the prediction of drug success in preventing CVD events. Future research should seek method-replication in diverse populations to determine the broad reprodicibility of these findings and potential clinical application.

Author Contributions: Conceptualization, S.M.A., C.C., S.E., and A.F.H.; methodology, S.M.A., S.E., A.F.H.; software, S.M.A.; resources, S.M.A.; writing–original draft preparation, S.M.A., T.N., H.F.; writing–review and editing, S.M.A., C.C., S.E., A.F.H.; supervision, S.M.A. All authors have read and agreed to the published version of the manuscript.

Funding: Vanderbilt University Medical Center's BioVU is supported by institutional funding and by the Vanderbilt CTSA grant UL1 TR000445 from NCATS/NIH. This study was supported by an NIH Pharmacogenomics Research Network (PGRN)—RIKEN Center for Integrative Medical Sciences (IMS) Global Alliance, and genome-wide genotyping was funded and performed by the IMS. This study was also supported by NHLBI/NIH grants U19 HL065962 and U01 HL069757. The authors of this study are also supported by a grant from the Andrew W. Mellon Foundation (SMA) and U5 4MD010723 (AFH).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Grundy, S.M.; Stone, N.J.; Bailey, A.L.; Beam, C.; Birtcher, K.K.; Blumenthal, R.S.; Braun, L.T.; de Ferranti, S.; Faiella-Tommasino, J.; Forman, D.E.; et al. 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA Guideline on the Management of Blood Cholesterol: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation* **2019**, *139*, e1082–e1143. [CrossRef] [PubMed]
2. Ramos, R.; García-Gil, M.; Comas-Cufí, M.; Quesada, M.; Marrugat, J.; Elosua, R.; Sala, J.; Grau, M.; Martí, R.; Ponjoan, A.; et al. Statins for Prevention of Cardiovascular Events in a Low-Risk Population With Low Ankle Brachial Index. *J. Am. Coll. Cardiol.* **2016**, *67*, 630–640. [CrossRef] [PubMed]
3. Gutierrez, J.; Ramirez, G.; Rundek, T.; Sacco, R.L. Statin therapy in the prevention of recurrent cardiovascular events: A sex-based meta-analysis. *Arch. Intern. Med.* **2012**, *172*, 909–919. [CrossRef] [PubMed]
4. Efficacy and safety of statin therapy in older people: A meta-analysis of individual participant data from 28 randomised controlled trials. *Lancet* **2019**, *393*, 407–415. [CrossRef]

5. Ramsey, L.B.; Johnson, S.G.; Caudle, K.E.; Haidar, C.E.; Voora, D.; Wilke, R.A.; Maxwell, W.D.; McLeod, H.L.; Krauss, R.M.; Roden, D.M.; et al. The clinical pharmacogenetics implementation consortium guideline for SLCO1B1 and simvastatin-induced myopathy: 2014 update. *Clin. Pharmacol. Ther.* **2014**, *96*, 423–428. [CrossRef] [PubMed]
6. Ruiz-Iruela, C.; Candás-Estébanez, B.; Pintó-Sala, X.; Baena-Díez, N.; Caixàs-Pedragós, A.; Güell-Miró, R.; Navarro-Badal, R.; Calmarza, P.; Puzo-Foncilla, J.L.; Alía-Ramos, P.; et al. Genetic contribution to lipid target achievement with statin therapy: A prospective study. *Pharm. J.* **2020**, *20*, 494–504. [CrossRef] [PubMed]
7. Kessler, T.; Vilne, B.; Schunkert, H. The impact of genome-wide association studies on the pathophysiology and therapy of cardiovascular disease. *EMBO Mol. Med.* **2016**, *8*, 688–701. [CrossRef] [PubMed]
8. Roguin, A.; Koch, W.; Kastrati, A.; Aronson, D.; Schomig, A.; Levy, A.P. Haptoglobin genotype is predictive of major adverse cardiac events in the 1-year period after percutaneous transluminal coronary angioplasty in individuals with diabetes. *Diabetes Care* **2003**, *26*, 2628–2631. [CrossRef] [PubMed]
9. Zhao, C.; Zhu, P.; Shen, Q.; Jin, L. Prospective association of a genetic risk score with major adverse cardiovascular events in patients with coronary artery disease. *Medicine* **2017**, *96*, e9473. [CrossRef]
10. Wang, L.; McLeod, H.L.; Weinshilboum, R.M. Genomics and drug response. *N. Engl. J. Med.* **2011**, *364*, 1144–1153. [CrossRef]
11. Gibson, G. On the utilization of polygenic risk scores for therapeutic targeting. *PLoS Genet.* **2019**, *15*, e1008060. [CrossRef] [PubMed]
12. Jiang, R.; Tang, W.; Wu, X.; Fu, W. A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinform.* **2009**, *10* (Suppl. S1), S65. [CrossRef] [PubMed]
13. Li, J.; Malley, J.D.; Andrew, A.S.; Karagas, M.R.; Moore, J.H. Detecting gene-gene interactions using a permutation-based random forest method. *BioData Min* **2016**, *9*, 14. [CrossRef] [PubMed]
14. Nembrini, S.; König, I.R.; Wright, M.N. The revival of the Gini importance? *Bioinformatics* **2018**, *34*, 3711–3718. [CrossRef] [PubMed]
15. Vailhé, B.; Vittet, D.; Feige, J.J. In vitro models of vasculogenesis and angiogenesis. *Lab. Investig.* **2001**, *81*, 439–452. [CrossRef]
16. Yang, W.; Ng, F.L.; Chan, K.; Pu, X.; Poston, R.N.; Ren, M.; An, W.; Zhang, R.; Wu, J.; Yan, S.; et al. Coronary-Heart-Disease-Associated Genetic Variant at the COL4A1/COL4A2 Locus Affects COL4A1/ COL4A2 Expression, Vascular Cell Survival, Atherosclerotic Plaque Stability and Risk of Myocardial Infarction. *PLoS Genet.* **2016**, *12*, e1006127. [CrossRef] [PubMed]
17. Vargas, J.D.; Manichaikul, A.; Wang, X.Q.; Rich, S.S.; Rotter, J.I.; Post, W.S.; Polak, J.F.; Budoff, M.J.; Bluemke, D.A. Common genetic variants and subclinical atherosclerosis: The Multi-Ethnic Study of Atherosclerosis (MESA). *Atherosclerosis* **2016**, *245*, 230–236. [CrossRef]
18. Dehghan, A.; Bis, J.C.; White, C.C.; Smith, A.V.; Morrison, A.C.; Cupples, L.A.; Trompet, S.; Chasman, D.I.; Lumley, T.; Völker, U.; et al. Genome-Wide Association Study for Incident Myocardial Infarction and Coronary Heart Disease in Prospective Cohort Studies: The CHARGE Consortium. *PLoS ONE* **2016**, *11*, e0144997. [CrossRef]
19. Vargas, J.D.; Manichaikul, A.; Wang, X.Q.; Rich, S.S.; Rotter, J.I.; Post, W.S.; Polak, J.F.; Budoff, M.J.; Bluemke, D.A. Detailed analysis of association between common single nucleotide polymorphisms and subclinical atherosclerosis: The Multi-ethnic Study of Atherosclerosis. *Data Brief* **2016**, *7*, 229–242. [CrossRef]
20. Lygirou, V.; Latosinska, A.; Makridakis, M.; Mullen, W.; Delles, C.; Schanstra, J.P.; Zoidakis, J.; Pieske, B.; Mischak, H.; Vlahou, A. Plasma proteomic analysis reveals altered protein abundances in cardiovascular disease. *J. Transl. Med.* **2018**, *16*, 104. [CrossRef]
21. Mesitskaya, D.F.; Syrkin, A.L.; Aksenova, M.G.; Zhang, Y.; Zamyatnin, A.A.; Kopylov, P.Y. Thromboxane A Synthase: A New Target for the Treatment of Cardiovascular Diseases. *Cardiovasc. Hematol. Agents Med. Chem.* **2018**, *16*, 81–87. [CrossRef]
22. Toumaniantz, G.; Ferland-McCollough, D.; Cario-Toumaniantz, C.; Pacaud, P.; Loirand, G. The Rho protein exchange factor Vav3 regulates vascular smooth muscle cell proliferation and migration. *Cardiovasc. Res.* **2010**, *86*, 131–140. [CrossRef]
23. Xu, J.Z.; Zhang, J.L.; Zhang, W.G. Antisense RNA: The new favorite in genetic research. *J. Zhejiang Univ. Sci. B* **2018**, *19*, 739–749. [CrossRef] [PubMed]

24. Aslibekyan, S.; Kabagambe, E.K.; Irvin, M.R.; Straka, R.J.; Borecki, I.B.; Tiwari, H.K.; Tsai, M.Y.; Hopkins, P.N.; Shen, J.; Lai, C.Q.; et al. A genome-wide association study of inflammatory biomarker changes in response to fenofibrate treatment in the Genetics of Lipid Lowering Drug and Diet Network. *Pharm. Genom.* **2012**, *22*, 191–197. [CrossRef]
25. Peeters, T.; Monteagudo, S.; Tylzanowski, P.; Luyten, F.P.; Lories, R.; Cailotto, F. SMOC2 inhibits calcification of osteoprogenitor and endothelial cells. *PLoS ONE* **2018**, *13*, e0198104. [CrossRef] [PubMed]
26. Howson, J.M.M.; Zhao, W.; Barnes, D.R.; Ho, W.K.; Young, R.; Paul, D.S.; Waite, L.L.; Freitag, D.F.; Fauman, E.B.; Salfati, E.L.; et al. Fifteen new risk loci for coronary artery disease highlight arterial-wall-specific mechanisms. *Nat. Genet.* **2017**, *49*, 1113–1119. [CrossRef] [PubMed]
27. Privratsky, J.R.; Paddock, C.M.; Florey, O.; Newman, D.K.; Muller, W.A.; Newman, P.J. Relative contribution of PECAM-1 adhesion and signaling to the maintenance of vascular integrity. *J. Cell Sci.* **2011**, *124*, 1477–1485. [CrossRef]
28. Ueland, T.; Åkerblom, A.; Ghukasyan, T.; Michelsen, A.E.; Becker, R.C.; Bertilsson, M.; Budaj, A.; Cornel, J.H.; Himmelmann, A.; James, S.K.; et al. ALCAM predicts future cardiovascular death in acute coronary syndromes: Insights from the PLATO trial. *Atherosclerosis* **2020**, *293*, 35–41. [CrossRef] [PubMed]
29. Shendre, A.; Irvin, M.R.; Wiener, H.; Zhi, D.; Limdi, N.A.; Overton, E.T.; Shrestha, S. Local Ancestry and Clinical Cardiovascular Events Among African Americans From the Atherosclerosis Risk in Communities Study. *J. Am. Heart Assoc.* **2017**, *6*. [CrossRef] [PubMed]
30. Folestad, E.; Kunath, A.; Wågsäter, D. PDGF-C and PDGF-D signaling in vascular diseases and animal models. *Mol. Aspects Med.* **2018**, *62*, 1–11. [CrossRef]
31. Moriya, J.; Wu, X.; Zavala-Solorio, J.; Ross, J.; Liang, X.H.; Ferrara, N. Platelet-derived growth factor C promotes revascularization in ischemic limbs of diabetic mice. *J. Vasc. Surg.* **2014**, *59*, 1402–1409. [CrossRef]
32. Zhang, H.; He, Y.; Dai, S.; Xu, Z.; Luo, Y.; Wan, T.; Luo, D.; Jones, D.; Tang, S.; Chen, H.; et al. AIP1 functions as an endogenous inhibitor of VEGFR2-mediated signaling and inflammatory angiogenesis in mice. *J. Clin. Invest.* **2008**, *118*, 3904–3916. [CrossRef]
33. Harrison, S.C.; Cooper, J.A.; Li, K.; Talmud, P.J.; Sofat, R.; Stephens, J.W.; Hamsten, A.; Sanders, J.; Montgomery, H.; Neil, A.; et al. Association of a sequence variant in DAB2IP with coronary heart disease. *Eur. Heart J.* **2012**, *33*, 881–888. [CrossRef] [PubMed]
34. Gretarsdottir, S.; Baas, A.F.; Thorleifsson, G.; Holm, H.; den Heijer, M.; de Vries, J.P.P.M.; Kranendonk, S.E.; Zeebregts, C.J.A.M.; van Sterkenburg, S.M.; Geelkerken, R.H.; et al. Genome-wide association study identifies a sequence variant within the DAB2IP gene conferring susceptibility to abdominal aortic aneurysm. *Nat. Genet.* **2010**, *42*, 692–697. [CrossRef]
35. Xu, J.J.; Jiang, L.; Xu, L.J.; Gao, Z.; Zhao, X.Y.; Zhang, Y.; Song, Y.; Liu, R.; Sun, K.; Gao, R.L.; et al. Association of CDKN2B-AS1 Polymorphisms with Premature Triple-vessel Coronary Disease and Their Sex Specificity in the Chinese Population. *Biomed. Environ. Sci.* **2018**, *31*, 787–796. [CrossRef]
36. Heit, C.; Jackson, B.C.; McAndrews, M.; Wright, M.W.; Thompson, D.C.; Silverman, G.A.; Nebert, D.W.; Vasiliou, V. Update of the human and mouse SERPIN gene superfamily. *Hum. Genomics* **2013**, *7*, 22. [CrossRef] [PubMed]
37. Aulchenko, Y.S.; Ripke, S.; Isaacs, A.; van Duijn, C.M. GenABEL: An R library for genome-wide association analysis. *Bioinformatics* **2007**, *23*, 1294–1296. [CrossRef]
38. Chang, C.C.; Chow, C.C.; Tellier, L.C.; Vattikuti, S.; Purcell, S.M.; Lee, J.J. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *Gigascience* **2015**, *4*, 7. [CrossRef] [PubMed]
39. Wright, M.; Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Stat. Softw. Artic.* **2017**, *77*, 1–17. [CrossRef]
40. Szymczak, S.; Holzinger, E.; Dasgupta, A.; Malley, J.D.; Molloy, A.M.; Mills, J.L.; Brody, L.C.; Stambolian, D.; Bailey-Wilson, J.E. r2VIM: A new variable selection method for random forests in genome-wide association studies. *BioData Min.* **2016**, *9*, 7. [CrossRef]
41. Degenhardt, F.; Seifert, S.; Szymczak, S. Evaluation of variable selection methods for random forests and omics data sets. *Brief. Bioinform.* **2019**, *20*, 492–503. [CrossRef]
42. Lewis Schmalohr, C.; Grossbach, J.; Clément-Ziza, M.; Beyer, A. Detection of epistatic interactions with Random Forest. *bioRxiv* **2018**. [CrossRef]
43. Berger, A. FUNDAMENTALS OF BIOSTATISTICS. *Am. J. Public Health Nat. Health* **1969**, *59*, 1266–1266. [CrossRef]

44. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. Royal Stat. Soc. Ser. B (Methodological)* **1995**, *57*, 289–300. [CrossRef]
45. Hothorn, T.; Hornik, K.; Zeileis, A. Unbiased Recursive Partitioning: A Conditional Inference Framework. *J. Comput. Graph. Stat.* **2006**, *15*, 651–674. [CrossRef]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.