Cis-epistasis at the LPA locus and risk of cardiovascular diseases

Lingyao Zeng†, Sylvain Moser¶, Nazanin Mirza-Schreiber‡, Claudia Lamina§, Oscar Franzen, Maria T. Kleber∥, Salome Mack, Till F. M. Andlauer¶, Beibejiang∥, Barbara Stiller†, Ling Li, Christina Willenborg, Matthews Munz, Thorsten Kessler, Adnan Kastrati, Karl-Ludwig Laugwitz, Jeanette Erdmann, Karsten Moebus, Markus M. Nöthen, Annette Peters, Konstantin Strauch, Christian Gieger, Thomas Meitinger, Elisabeth Steinhausen-Thiessen, Winfried März, Andres Metspalu, Johan L. M. Björkergren, Nilesh J. Samani, Florian Kronenberg, Bertram Müller-Myhsok, and Heribert Schunkert.

Aims

Coronary artery disease (CAD) has a strong genetic predisposition. However, despite substantial discoveries made by genome-wide association studies (GWAS), a large proportion of heritability awaits identification. Non-additive genetic effects might be responsible for part of the unaccounted genetic variance. Here, we attempted a proof-of-concept study to identify non-additive genetic effects, namely epistatic interactions, associated with CAD.

Methods and results

We tested for epistatic interactions in 10 CAD case-control studies and UK Biobank with focus on 8068 SNPs at 56 loci with known associations with CAD risk. We identified a SNP pair located in cis at the LPA locus and risk of LPA.

†The first three authors are shared first authors.
¶Corresponding authors. Tel: +49 89/1218-4037; fax: +49 89/1218-4013; E-mail: schunkert@dhm.mhn.de (H.S.); Tel: +49 89/30622-246; fax: +49 89/30622-610, E-mail: bm@psch.mpg.de (B-M.M.)
© The Author(s) 2021. Published by Oxford University Press on behalf of the European Society of Cardiology.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
arterial disease (OR = 1.22, \( P = 2.32 \times 10^{-4} \)), aortic stenosis (OR = 1.47, \( P = 6.95 \times 10^{-7} \)), hepatic lipoprotein(a) (Lp(a)) transcript levels (beta = 0.39, \( P = 1.41 \times 10^{-9} \)), and Lp(a) serum levels (beta = 0.58, \( P = 8.7 \times 10^{-32} \)), while individual SNPs displayed no association. Further exploration of the LPA locus revealed a strong dependency of these associations on a rare variant, rs140570886, that was previously associated with Lp(a) levels. We confirmed increased CAD risk for heterozygous (relative OR = 1.46, \( P = 9.97 \times 10^{-32} \)) and individuals homozygous for the minor allele (relative OR = 1.77, \( P = 0.09 \)) of rs140570886. Using forward model selection, we also show that epistatic interactions between rs140570886, rs9458001, and rs1800769 modulate the effects of the rs140570886 risk allele.

### Conclusions

These results demonstrate the feasibility of a large-scale knowledge-based epistasis scan and provide rare evidence of an epistatic interaction in a complex human disease. We were directed to a variant (rs140570886) influencing risk through additive genetic as well as epistatic effects. In summary, this study provides deeper insights into the genetic architecture of a locus important for cardiovascular diseases.

### Graphical Abstract

A decade of genome-wide association studies (GWAS) shed light on the genetic architecture of the disease, discovering 163 genetic loci associated with CAD risk.\(^5\) About a quarter of CAD heritability can be explained by additive effects of these and other common genetic variants.\(^6\) More complex models involving gene regulatory networks\(^4\) may help to better explain the heritability of the disease. In addition, at some of these loci, multiple independent signals were described, showing intra-locus allelic heterogeneity.\(^7\) Until now, non-additive genetic effects,
such as epistatic interactions, are largely neglected for explaining the heritability of CAD. However, epistasis has been postulated by some to account for part of this ‘missing heritability’ and has also been found to act alongside additive effects to influence complex phenotypes.

Epistatic interactions have profound effects in bacteria as well as in other higher model organisms and have been shown to regulate some quantitative traits in humans. However, evidence of epistasis in human genetics remains very scarce, because individual-level data with large sample sizes are required for epistasis studies. Moreover, the combinatorial nature of epistasis makes hypothesis-free genome-wide interaction analyses (GWIA) computationally demanding and plagued with a high multiple testing burden. Finally, associations based on interactions appear to suffer from a low replication rate, and genetic interactions are sometimes difficult to disentangle from the tagging of haplotypes. Indeed, a non-causal combination of alleles at multiple SNPs co-inherited with a rare causal variant could act as a tag for this variant.

To face the computational complexities in search for interacting loci affecting risk for CAD, we conducted a two-stage statistical scanning procedure for epistasis using a GPU-accelerated software on individual-level data from several GWAS on CAD. The scan was based on susceptibility regions defined around the top 56 known CAD loci, thereby limiting the multiple testing correction burden while maximizing the likelihood to discover biologically relevant interactions.

2 Methods

2.1 Cohorts

2.1.1 CAD case–control studies

Individual-level genotypes were obtained from 10 CAD case–control studies. From Germany: the German Myocardial Infarction Family Studies (GerMIFS) I, II, III (KORA), the LuDiQwahsen Risk and Cardiovascular Health Study (LURIC)23, from Germany, England, and France: Cardiogenics; from England: Wellcome Trust Case Control Consortium (WTCCC)24,25; from France, Italy, Germany, and the USA: Myocardial Infarction Family Studies and STARNET-Study.

2.1.2 UK Biobank

The UK Biobank (UKBB) project (http://www.ukbiobank.ac.uk) is a large prospective cohort study of ~500,000 individuals from across the UK, aged 40–69 years at recruitment. In the present study, CAD cases were defined using the ‘SOFT’ and ‘HARD’ criteria, i.e., as individuals with fatal or non-fatal myocardial infarction (MI), percutaneous transluminal coronary angioplasty (PTCA), coronary artery bypass grafting (CABG), chronic ischemic heart disease (IHD), and angina. Peripheral arterial disease (PAD) cases were defined as self-reported history of aortic stenosis or either hospitalization or death due to ICD9-424.1 or ICD10-I35.0. The post-imputation sample quality control (QC) performed in the UKBB dataset is detailed in the Supplementary material online. Methods. UKBB data were accessed under the approval of UKBB within project 9922. The study was conducted following the principles of the declaration of Helsinki and all participants gave prior written informed consent.

2.1.3 KORA F3/F4 studies and STARNET-Study

Individual-level genotypes were obtained from population studies from Augsburg, Germany,29 KORA F3 and KORA F430,31 along with lipid measurements including total lipoprotein(a) (Lp(a)) levels and the number of Kringle repeats of the Lp(a) protein. RNAseq data were generated from liver tissue of 522 CAGB CAD patients from the Stockholm-Tartu Reverse Network Engineering Task (STARNET) study.32 These studies obtained institutional review board approval from their local Ethical Committees and were performed in accordance with the 1964 Helsinki Declaration and its later amendments. All participants gave prior written informed consent. Further information about these studies is provided in the Supplementary material online, Methods.

2.2 Epistasis scan

2.2.1 Broad sense CAD susceptibility region

We focused our analysis on 56 loci with previous evidence from GWAS on CAD30,35 (Supplementary material online, Table S14) in order to restrict the number of variants for testing of statistical epistasis. Our aim was to enhance computation time and the likelihood of true positive findings by easing the multiple testing correction burden. CAD susceptibility regions were defined as ±500 kb around each of the 56 lead SNPs. This window size was chosen to capture the loci as completely as possible while minimizing the computational burden: the variance explained by the lead SNPs accounted for only 46% of the variance explained when including their flanking ±500 kb regions (Supplementary material online, Figure S2 and Methods). We then pruned the variants in each region to 8,068 SNPs with pairwise $r^2 < 0.5$ located in the broad CAD susceptibility regions.

2.2.2 Statistical interaction analysis

We used the general framework for detecting statistical epistasis in quantitative genetics as proposed by Hansen and Wagner on the pairwise epistasis between two loci (SNPs) and implemented a two-stage statistical scanning procedure (Figure 1). The first step of the testing procedure consisted in a loose but fast statistical filtering using the GLIDE GPU computation tool. For each possible pair of SNPs, we fitted a linear model with the CAD phenotype as the dependent variable and the marginal effect of the two SNPs and their interaction term as predictors. (Eq. (1)). Each SNP’s genotype was encoded in four different models, dosage, dominant, recessive, and heterozygous with respect to the minor allele and all combinations of $2^4$ models were tested for each pair of SNPs.

$$y \sim b_0 + b_1 \times \text{SNP}_1 + b_2 \times \text{SNP}_2 + b_{12} \times \text{SNP}_1 \times \text{SNP}_2$$

A relatively loose and arbitrary significance level ($P < 1 \times 10^{-6}$) was applied for primary filtering, with the assumption that if true epistasis existed between two SNPs, signals of moderate strength should be detectable between the SNPs within the corresponding linkage disequilibrium (LD) block. This threshold was defined with the aim to detect such pair in LD with the true epistasis signal and to forward a manageable number of pairs to the second step.

The second step included the fine-mapping of candidate SNP pairs to screen for the strongest signal among the SNPs in the same LD block. For this purpose, we used R to fit a logistic regression model, slower than the linear model used in step 1, but suited better for the binary CAD phenotype, and extended Eq. (1) to correct for population structure by adding the first 10 multi-dimensional scaling (MDS) components of the genetic relationship
In the discovery phase, the same epistasis testing procedure was performed in each of the 10 CAD case-control studies separately. The models used genotype data imputed to the 1000 Genomes Phase 3 (1000GP3) reference panel. This regression analysis was followed by fixed-effects meta-analysis to estimate the overall effect size and standard error. The final epistasis pair of interest was then re-analysed in the same studies imputed using the Haplotype Reference Consortium (HRC) reference panel, to enable a more complete coverage of the region of interest in all 10 cohorts. Thereafter, this imputation based on the larger HRC reference was used for the remainder of the manuscript.

2.2.3 Prioritizing candidate SNP pairs of epistasis of CAD

After the detection of SNP pairs showing statistically significant epistatic effects on the risk for CAD, we prioritized candidate pairs based on the following four criteria:

1. We retained only SNP pairs with a high replication potential [i.e. displaying statistical epistasis both significantly ($P < 4.6 \times 10^{-9}$) and consistently (effect sizes pointing in the same direction) in at least eight of the 10 studies in the discovery data, based on both imputations].
2. LD between two target SNPs located on the same chromosome $r^2 < 0.2$.
3. Weak interaction signals detectable between SNPs that show an LD $r^2 > 0.5$ with any of the two interacting SNPs.
4. The effect of the interaction term is independent (i.e. $P$-value in conditional models <7.8 $\times 10^{-6}$) of any available third variant in conditional analyses.

2.3 Conditional analysis

The aim of the conditional analysis was to test whether the statistical epistasis effects were independent from a third SNP. To this end, we tested for the independence of the interaction term against the SNPs located within a ± 200kb window around the epistatic loci and any known CAD GWAS SNPs that survived the original QC procedure. This window size was chosen to capture all SNPs in significant LD with the pair of interest. Indeed, it has been shown that LD decay with physical distance and is close to 0 at 200 kb. For each of these SNPs, we used R to compute a likelihood ratio test (LRT) between a model including the additive effect of the two target SNPs and the additive effect of the conditioning SNP (all coded as minor allele dosages) [Eq. (3)] and a model including the interaction term in addition [Eq. (4)].

$y = b_0 + b_1 \times snp_1 + b_2 \times snp_2 + b_3 \times snp_1 \times snp_2 + b_4 \times MDS_1 + b_5 \times MDS_2 + \ldots + b_{10} \times MDS_{10}$

(2)

$y = b_0 + b_1 \times snp_1 + b_2 \times snp_2 + b_3 \times snp_1 \times snp_2 + b_4 \times cov_1 + b_5 \times cov_2 + \ldots + b_{10} \times cov_{10}$

(3)

(4)
2.4 Relative effect sizes and analyses of intermediate traits

Genotypic effect sizes for the different rs140570886 genotypes were computed by regression analysis in R using the dosage genetic model. Association analysis for the continuous intermediate traits Lp(a) protein levels, LPA mRNA levels, and KIV repeats were also performed using linear regression. Lp(a) protein levels were highly skewed and Inverse Normal Transformation was applied prior association. The relative effect size for the three-SNP haplotypes was computed via haplotype estimation followed by fitting a generalized linear model with the R package haplo.stats. More detailed descriptions of these statistical procedures are provided in the Supplementary material online, Methods.

3 Results

3.1 Discovery of SNP pairs associated with CAD risk

We identified 56 previously known CAD risk loci from two previous GWAS (Supplementary material online, Table S14). For our study, we extracted 8068 LD-independent candidate variants within 500 kb of the respective lead SNPs. We observed that these extended regions explained more phenotypic variance than the respective lead SNPs alone (Supplementary material online, Figure S2 and Methods). Testing for statistical interactions was carried out on all pairwise SNPs along with a two-step scheme (described in Figure 1) on imputed genotypes from 29,755 participants of 10 European CAD case–control studies (17–22,24–26 (Figure 2). Four SNP pairs displayed consistent (i.e. in at least 8 of 10 studies) and significant (i.e. \( P < 4.618 \times 10^{-3} \)) effects and thus met our criteria as candidates for epistasis (Supplementary material online, Table S2). Among these four pairs, two (rs1800769 \( \times \) rs9458001 and rs116632378 \( \times \) rs3798220) replicated in the UKBB. The top SNP pair (rs1800769 \( \times \) rs9458001) showed the strongest effect in a dosage–dosage model and was prioritized for further investigation (Supplementary material online, Table S3).

Both rs1800769 and rs9458001 map to chromosome 6, close to the LPA locus (Figure 3B), and are not in LD with each other \( (r^2 = 0.014, D = 0.533, \text{Table } 1) \). None of the SNPs were associated with CAD risk by itself in an additive model \( (P = 0.59, \text{odds ratio (OR)} = 0.99 \text{ for rs1800769}[T]; \ P = 0.08, \text{OR} = 1.04 \text{ for rs9458001}[A]) \). However, the interaction term displayed a strong association \( (\text{OR}_{\text{int}} = 1.42, P = 1.75 \times 10^{-13} \text{ for the rs1800769}[T] \times \text{rs9458001}[A] \text{ interaction term}) \). In this case, as both SNP were encoded in the alternate genetic model, the OR can be interpreted as the increase in likelihood to suffer from CAD associated with an increase of one unit in the product between the number of minor alleles at each of the interacting SNPs. The results were reproduced in the same dataset imputed with the HRC reference panel \( ^{26} \) using rs1652507 (LD with rs1800769, \( r^2 = 0.965, D^2 = 0.991, \text{Table } 1) \) as a proxy for rs1800769 (OR = 0.98, P = 0.38 for rs1652507[C]; OR = 1.03, P = 0.1 for rs9458001[A], and OR_int = 1.36, P = 1.07 \times 10^{-11} \text{ for the rs1652507[C] \times rs9458001[A] interaction term}) (Supplementary material online, Table S6). Thus, the results were qualitatively independent of the imputation panel. This newer and denser imputation with this proxy variant was used for the remainder of the manuscript.

3.2 Replication and association with further traits

The UKBB dataset (controls/cases \( n = 285 \times 26792 \)), used as an external replication sample (Figure 2), showed a consistent interaction effect of this SNP pair for CAD \( (\text{OR}_{\text{int}} = 1.15, P = 5.67 \times 10^{-10} \text{ for the rs1652507[C] \times rs9458001[A] interaction term}) \) in the dosage model (Supplementary material online, Table S8). Moreover, we found interaction effects in the same direction and with a comparable magnitude on peripheral vascular disease (controls/cases \( n = 475 \times 4460, \text{OR}_{\text{int}} = 1.22, P = 2.32 \times 10^{-4} \) and aortic valve stenosis (controls/cases \( n = 477 \times 2023, \text{OR}_{\text{int}} = 1.47, P = 6.95 \times 10^{-3} \)) (Supplementary material online, Table S8), conditions known to be affected by Lp(a) plasma levels. \( ^{37,38} \)

Next, we analysed the influence of the interaction term rs1800769 \( \times \) rs9458001 on circulating Lp(a) levels in a German population-based study \((\text{KORA F3/F4,} \ n = 5953) \) (Figure 2). In addition to the association of each SNP separately, we identified a strong interaction effect of both SNPs on inverse-rank normal-transformed (INT) Lp(a) levels \( (\beta = 0.58, P = 8.7 \times 10^{-32} \text{ with the SNP encoded in the dosage model}) \) (Supplementary material online, Table S6). In the LURIC study, we replicated the significant statistical interaction for INT Lp(a) levels \( (\beta = 0.56, P = 6.93 \times 10^{-14}) \) and found no other circulating factor displaying such effects (data not shown).

Finally, we extended our investigation to LPA mRNA expression in liver tissue (Section 2, STARNET study, \( n = 522 \)) (Figure 2), where LPA is transcribed into Apo(a) and further assembled with an LDL-like particle into Lp(a). A significant interaction between the two SNPs was found \( (P = 1.4 \times 10^{-6}) \) and the effects on LPA mRNA expression correlated with the circulating Lp(a) levels measured in KORA F3/F4 for various genotype subgroups (Supplementary material online, Table S6), suggesting that differential gene expression activity underlies a large component of statistical interaction related to the two SNPs.

3.3 Rs140570886-related effects at the LPA locus

An inherent challenge in testing for epistasis of nearby SNPs, even if they are in very low LD, is to discriminate interacting SNPs from SNPs representing a specific haplotype. In order to explore the latter possibility, we assessed the interaction effect after conditioning for any known susceptibility SNPs for CAD \( (n = 158, \text{Supplementary material online, Table S4}) \) or any available SNP in the flanking \( \pm 200 \text{ kb region}) \). The LPA region conditional analysis (see Section 2) did not yield any significant results (Supplementary material online, Table S5). However, studying GWAS lead SNPs (Supplementary material online, Table S4) uncovered that rs3798220 reduced the significance of the rs1652507 \( \times \) rs9458001 interaction term \( (\text{increase from } P = 1.07 \times 10^{-11} \text{ to } P = 2.08 \times 10^{-5} \text{, likelihood ratio test}) \).

In follow-up analyses, we also analysed the influence of the rare variant rs140570886 at the LPA locus, previously shown to be univariately
associated with Lp(a) levels. This variant was not included in our primary analysis because its minor allele frequency was lower than our QC threshold (see Section 2), but was pointed out to us as requiring special attention. We therefore specifically investigated rs140570886 in the conditional analysis and observed a drastic decrease in the statistical support for the rs1652507/C2 rs9458001 interaction term (from $P = 8.95 \times 10^{-14}$ to $P = 0.022$, likelihood ratio test). In order to test if these two SNPs (LD between rs140570886 and rs3798220: $r^2 = 0.808$, $D' = 0.899$) represented independent signals, we performed model selection using the likelihood ratio test. Adding rs3798220 to a model already containing rs140570886 did not improve the fit significantly ($P = 0.49$, likelihood ratio test). We therefore conclude that rs3798220 is not independent of rs140570886 and did not assess this SNP in further analyses.

We next investigated the additive effect of rs140570886 on CAD risk and found a significant association (OR = 1.98, $P = 1.14 \times 10^{-21}$, Figure 3A, Supplementary material online, Table S10). We replicated this association in the UKBB dataset (OR = 1.46, $P = 2.77 \times 10^{-32}$) (Figure 3A). Furthermore, in the UKBB, we found an association in the same
direction and comparable magnitude for peripheral arterial disease (controls/cases: \(n = 315,072/29,877\), OR = 1.43, \(P = 7.83 \times 10^{-7}\)) and aortic valve stenosis (controls/cases: \(n = 315,072/29,877\), OR = 1.71, \(P = 1.25 \times 10^{-7}\)) (Supplementary material online, Table S10), both of which are manifestations of atherosclerosis in coronary arteries for which \(\text{Lp(a)}\) plasma levels affect risk.37,38

To assess the contribution of rs140570886 genotypes to disease risk beyond the additive model, we next computed genotypic ORs for heterozygous [T/C] and minor allele homozygous genotypes [C/C] compared to the major allele homozygous reference genotype [T/T]. The genotypic model has the advantage that it does not make any assumption on the underlying genetic model. In the meta-analysis of the 10 CAD studies, we observed an OR of 1.88 (\(P = 2.32 \times 10^{-18}\)) for the T/C heterozygous genotype (Figure 4A, Supplementary material online, Table S10). A reliable effect estimate could not be calculated for the minor allele homozygous genotype C/C, due to its low frequency. The result for the T/C genotype was replicated in UKBB (OR = 1.46, \(P = 9.97 \times 10^{-32}\)) where we observed a trend for a higher relative OR for CC-homozygous subjects, although this was non-significant, likely due to its low frequency (OR = 1.77, \(P = 0.09\); Figure 4B). This increase of the genotypic OR with the number of minor alleles suggests that the additive genetic model is indeed likely correct for rs140570886. Coherent with this, the saturated genotypic model does not provide a better fit than the additive model (\(P = 0.12\), likelihood ratio test).

Values in Italic were computed in the European sub-samples of the 1000 Genomes Project using the LDmatrix tool (https://ldlink.nci.nih.gov/) as rs1800769 was absent from the HRC imputation panel. Other values were computed in each of the 10 CAD studies separately and averaged.

### Table 1 Linkage disequilibrium and minor allele frequency

| SNP        | rs3798220 | rs140570886 | rs1652507 | rs1800769 | rs9458001 |
|------------|-----------|-------------|-----------|-----------|-----------|
| rs3798220  | 0.01932745| 0.703204    | 0.05355715| NA        | 0.0273448 |
| rs140570886| 0.703204  | 0.01551685  | 0.0759549 | 0.1591815 | 0.0177401 |
| rs1652507  | 0.05355715| 0.0759549   | 0.965     | 0.096     | 0.014     |
| rs1800769  | NA        | NA          | 0.965     | 0.014     |           |
| rs9458001  | 0.0273448 | 0.0393634   | 0.0177401 | 0.014     | 0.2250345 |

This table shows the pairwise \(r^2\) measure of Linkage Disequilibrium (LD) between the reported SNPs and their respective minor allele frequency (MAF) in bold on the diagonal. Values in Italic were computed in the European sub-samples of the 1000 Genomes Project using the LDmatrix tool (https://ldlink.nci.nih.gov/) as rs1800769 was absent from the HRC imputation panel. Other values were computed in each of the 10 CAD studies separately and averaged.
Circulating Lp(a) levels are modulated by at least two independent mechanisms. First, they are inversely correlated with the number of Kringle IV type 2 repeats (KIV-2 CNV), with fewer KIV-2 CNV repeats associated with more Lp(a) release from liver cells. They account for about 18% of the variability in Lp(a) levels in Western Europeans. However, individuals with the same number of KIV-2 CNV repeats may differ up to 200-fold with respect to their Lp(a) levels, suggesting transcriptional mechanisms. In the KORA cohorts, we observed an association of rs140570886 with the KIV-2 CNV, with heterozygous rs140570886 carriers having fewer KIV-2 CNV repeats (beta = -5.74, P = 3.55 x 10^{-26}) (Supplementary material online, Table S10 and Methods). However, rs140570886 was in minimal LD with the reported 61 KIV-2 CNV-representing variants and the three independent modifier variants that influence the relationship between KIV-2 CNV and Lp(a) cholesterol (data not shown). More importantly, the effect of rs140570886 on Lp(a) levels remained highly significant after adjustment for the KIV-2 CNV (beta = 1.11, P = 1.94 x 10^{-34}) (Figure 4C, Supplementary material online, Table S10). This strongly suggests that the effect of rs140570886 on Lp(a) levels is independent of the KIV-2 CNV and might therefore be modulated by transcriptional regulation. In accordance with this hypothesis, we found rs140570886 to be part of a significant expression quantitative trait locus (eQTL) with LPA mRNA expression levels in liver tissue, where LPA is transcribed to Apo(a) and further assembled into Lp(a) (GTEx V8, normalized effect size = 0.98, P = 1.2 x 10^{-7}).

### 3.4 Interaction between rs140570886 and the rs1652507/rs9458001 pair

Although it appeared that part of the rs1652507 x rs9458001 interaction was due to tagging of an rs140570886-related effect, we wondered if epistasis could still be present. To investigate this possibility, we applied a likelihood ratio tests-based forward model selection procedure starting with only rs140570886 going up to a model including all main effects and interactions between the four SNPs, rs1652507, rs9458001, rs140570886, and rs3798220 (Supplementary material online, Methods). To increase statistical power, we here analysed the CAD studies and UKBB jointly. We observed a significant increase in model fit when rs1652507 and rs9458001 were added as predictors to the model already containing rs140570886 (Table 2). The addition of the rs1652507 x rs9458001 interaction term to this second model did not improve the fit further, coherent with the observed drop in the significance when conditioning the model containing the interaction term on rs140570886. However, the model fit increased significantly and reached its best level when all two-way and three-way interactions were added to the model. The direct comparison of this two-and-three-way interaction model to the rs140570886-only model yielded a P-value of the same magnitude as the original P-value threshold used for the epistasis screening (P = 6.46 x 10^{-7}). Using a type-III Sum of Squares ANOVA to dissect this final model, provided further insights into the importance of the different coefficients (Supplementary material online, Table S12): we observed in the two-and-three-way interactions model that the additive effect of rs140570886 became non-significant while the additive effect of rs1652507 reached significance. Moreover, albeit the originally discovered rs1652507 x rs9458001 interaction term became non-significant, we observed nominally significant interactions of both rs1652507 and rs9458001 with rs140570886. These results suggest that the additive effect of rs140570886 on CAD risk might actually be caused by more complicated patterns of cis-epistatic interactions.

To better understand the genetics underlying this statistical model, we computed the relative OR for each of the eight possible haplotypes. It appeared that all haplotypes including the major T allele for rs140570886 showed similar ORs. Interestingly, we observed that the effect size varied profoundly across haplotypes containing the rs140570886 minor allele C, depending on the rs1652507 genotype (red vs. blue on Figure 5, Supplementary material online, Table S13). Moreover, for the haplotype rs140570886[C]—rs1652507[T], we observed that the ORs were much lower for the [A] as compared to the [G] allele at rs9458001, although the standard errors were large due to the low frequencies of rarer haplotypes (Figure 5, Supplementary material online, Table S13). These two observations reflect the marginally significant interaction coefficients between rs140570886 and rs1652507 and between rs140570886 and rs9458001 in the two- and three-way interaction model (Supplementary material online, Table S12).

Finally, in a further attempt to distinguish epistatic interactions involving these three SNPs from a haplotype effect, we compared different models containing SNPs encoded in the additive model and either haplotypes, interactions, or both, using the Akaike Information Criterion (AIC) (Table 3). The model including SNPs and their interactions but no

### Table 2 ANOVA table reporting likelihood ratio test results for nested model in the model selection procedure

| Model | Residuals. Df | Residuals deviance | Df | Deviance | P-value | P-value LRT model 2 |
|-------|---------------|--------------------|----|----------|---------|---------------------|
| (1) CAD ~ covariates | 342 046 | 222 021 | NA | NA | NA | NA |
| (2) CAD ~ rs140570886 + covariates | 342 045 | 221 804 | 1 | 217.04 | 4 x 10^{-49} | NA |
| (3) CAD ~ rs140570886 + rs9458005 + rs1652507 + covariates | 342 043 | 221 770 | 2 | 33.57 | 5.1 x 10^{-8} | 5.1 x 10^{-8} |
| (4) CAD ~ rs140570886 + rs9458005 + rs1652507 + covariates | 342 042 | 221 768 | 1 | 2.32 | 0.13 | 7.9 x 10^{-6} |
| (5) CAD ~ rs140570886 * rs9458005 * rs1652507 + covariates | 342 039 | 221 755 | 3 | 13.42 | 0.004 | 6.5 x 10^{-9} |
| (6) CAD ~ rs140570886 + rs9458005 + rs1652507 + rs3798220 + covariates | 342 038 | 221 753 | 1 | 1.93 | 0.16 | 8.2 x 10^{-9} |
| (7) CAD ~ rs140570886 * rs9458005 * rs1652507 * rs3798220 + covariates | 342 031 | 221 746 | 7 | 6.22 | 0.51 | 8.2 x 10^{-9} |

The table displays the result of a series of successive likelihood ratio test between a nested model of increasing complexity performed on the merged dataset including the 10 CAD studies and the UK Biobank dataset. The first and second columns report the Residual Deviance and degrees of freedom from each row’s model. The ‘DF’ and ‘Deviance’ columns respectively report the difference in degrees of freedom and deviance between each row’s model and the model from the previous row. The P-value column reports the P-value of the likelihood ratio test between each row’s model and the previous one. The P-value LRT model 2 column reports the P-value of the likelihood ratio test between each model and the model containing only the additive effect of rs140570886. The * operator denotes factor crossing: a*b is interpreted as a + b + a x b. The 10 multi-dimensional scaling components of the genotype variance and the study were included as covariates in every model. Tables showing the results of the same analysis with 3, 5, or 7 MDS components are provided in the Supplementary material online, Tables S15–S17.
haplotypes showed the best AIC. This result firstly confirms that interactions between the three SNPs improve model fit compared to an additive only model. Secondly, it suggests the presence of a real epistatic interaction between the three SNPs rather than an exclusive haplotype effect. The likelihood ratio test applied to the nested model confirmed this interpretation. Indeed, the most complex model, including SNPs, haplotypes, and interactions, did not provide a better fit than the one without the haplotypes ($P = 0.26$, likelihood ratio test), whereas adding...
interactions to the SNPs model improved the fit significantly \( (P = 0.0034, \text{likelihood ratio test}) \). Although we cannot exclude the involvement of other rare or non-typed variants and lack the statistical power necessary for these two interactions to reach the significance threshold predefined for the scan, these results demonstrate a complex genetic architecture involving non-additive and likely epistatic effects in the LPA region, underlying the regulation of Lp(a) expression and CAD risk.

4 Discussion

We report a two-stage testing procedure for epistatic interactions affecting CAD risk. Our analysis identified two SNPs at the LPA locus that individually had no effect but jointly displayed a strong statistical association with expression of LPA mRNA in liver, Lp(a) levels in serum, and with risk for CAD, peripheral arterial disease, and aortic stenosis. Further exploration of the locus revealed that parts of these associations were explained by tagging of a low-frequency variant (rs140570886), which, in parallel with our study, was found to be associated with Lp(a) levels. In addition, we detected a complex pattern of interactions between this variant and two other SNPs in the LPA region. Together, these findings firstly provide evidence of epistatic interaction in a complex human disease and provide deeper insights into the genetic architecture of an important locus for cardiovascular risk. At the same time, these data highlight the challenges in confirming epistatic interactions affecting disease risk in humans.

We focused our search for pairwise epistatic interactions on 8068 SNPs at 56 regions that had been found to be associated at genome-wide significance with CAD. Indeed, the selected

![Figure 5](https://academic.oup.com/cardiovascres/article/118/4/1088/6240984) Relative effect of the rs140570886-rs1652507-rs9458001 3-SNP haplotypes on CAD risk. Relative odds ratio (OR; with reference to the most frequent TTG haplotypes) for the eight possible 3-SNP haplotypes on CAD risk. The red and blue colours represent the base at the rs1652507 SNP, the square and triangle shapes represent the base at the rs9458001 SNP and the position on the X-axis represent the base at the rs140570886 SNP. Together they indicate one of the eight possible 3-SNP haplotypes. The putative haplotypes were computed using the hapasoc R package on a merged dataset of the 10 CAD studies \( (N = 29755) \). Error bars represent the standard error of the log OR.
Table 3 Model selection using AIC and Likelihood ratio test confirms epistatic interactions at the LPA locus

| Model                          | Model name | AIC      | Comparison M_SNPs | Comparison M_interact |
|--------------------------------|------------|----------|-------------------|-----------------------|
| No genetics                    | M_null     | 222 063.0| NA                | NA                    |
| Haplotypes                     | M_haplo    | 221 813.6| NA                | NA                    |
| SNPs                           | M_SNPs     | 221 818.3| NA                | NA                    |
| SNPs + interactions            | M_interact | 221 810.6| 0.0034            | NA                    |
| Haplotypes + SNPs + interactions| M_full    | 221 814.4| 0.0268            | 0.262                 |

This table displays the Akaike Information Criterion (AIC) and results of likelihood ratio test for nested models of increasing complexity performed on the merged dataset including the 10 CAD studies and the UK Biobank dataset. The ‘Comparison M_SNPs’ and ‘Comparison M_interact’ columns respectively report the P-values of the likelihood ratio tests with the M_SNPs and M_interact models as null model. The 10 multi-dimensional scaling components of the genetic variance and were included as covariates in every model. NA, non-applicable.

window of LD-pruned SNPs contained two-fold more information on CAD heritability than the respective lead SNPs. Nevertheless, we found only four potentially interacting SNP pairs, which highlights the challenge to identify true epistasis modulating a human trait. The top-ranking interacting pair was located in cis at the LPA locus. Conditional analyses, aiming to determine the independence of the epistatic signal between rs1652507 and rs9458001 from other neighbouring SNPs, revealed a strong dependence on rs140570886. Thus, the seemingly strongest epistatic SNP pair tagged a rare genotype with profound effects on the phenotype. Further investigation of this variant showed its strong association with CAD and Lp(a) protein levels. rs140570886 has been previously associated with cardiovascular disease (CVD) using a new integrative framework named FiOIODR.45 Our result, thus, using a well-established logistic regression model,45 confirmed the association of rs140570886 with diseases of the cardiovascular system. We, moreover, replicated the association of rs140570886 with Lp(a) levels reported by Mack et al.19 and identified data collections that indicate an association of rs140570886 with CAD.7 In addition, we report rs140570886 effects on Lp(a) levels to be independent of the KIV CNV repeats and to be a strong eQTL for the KIV CNV repeats and to be a strong eQTL for the LPA gene expression. Taken together, these findings support the hypothesis that rs140570886 mediates CAD risk through the Lp(a) levels via transcriptional regulation.

An important methodological point highlighted by this study is the importance of the conditional follow-up analyses in the investigation of epistatic interactions. Indeed, an inherent challenge in testing for epistasis of nearby SNPs, even if they are in very low LD, is to discriminate truly interacting SNPs from SNPs tagging a specific haplotype.46 Resolving the dependence structure at the epistatic locus, by conditioning the interaction effect on the neighbouring SNPs, allowed us to simultaneously identify a tagged rarer variant and to fine-map the epistatic interaction at the LPA locus.

The combinatorial nature of interactions has been a major hold-up in epistasis testing because it leads to an enormous search space and a high multiple testing correction burden.15,47 Methods to reduce this space can be divided into two categories: data-driven and knowledge-driven methods.48 We applied a data-driven approach in the present study, focusing on previously associated loci, for two reasons. First, variants already shown to be linked to the disease are likely to be functionally important. Second, if epistatic effects were detected among such variants, these effects would be more likely to affect the condition. Since regulatory variants might be located in the flanking region of the prioritized loci, we extended the search space to these regions. The discovery of four pairs of interacting SNPs using this filtering approach demonstrates its advantage over a hypothesis-free approach, in which these pairs would not have reached statistical significance due to having to correct for more tests.

The findings relative to the genetic architecture of the LPA locus reported in this study carry a special clinical relevance for CAD risk detection and treatment. Indeed, Lp(a) concentrations have been shown to be usable for CAD risk prediction. For example, the Copenhagen City Heart Study showed for individuals above the 95th percentiles of the Lp(a) concentration to have a 2.5-fold higher CAD risk compared to individuals in the lowest quartile.37 Although Lp(a) concentration measurement and isoform determination are sufficient assays to estimate CAD risk encoded at the LPA locus,49 polygenic risk scores might play an additional role in the assessment of CAD risk in the future.5 Indeed, with the rapid drop of genotyping cost, individual genotype data are becoming a basic component of biobanks and clinical settings.50 With this perspective, a better understanding of the genetic architecture of the LPA locus and the incorporation of non-additive genetic effects, such as those reported in this study, might enhance the predictive power of polygenic risk scores and help the development of individually tailored disease prevention,51 which, in the future, may involve a pharmacological Lp(a) reduction.52

While a single epistatic interaction as reported in this manuscript is very unlikely to improve risk prediction on its own compared to polygenic risk score based on millions of SNPs, numerous interactions—if identified—might do so. Indeed, several observations argue in this direction. First, simulations and analyses by others indicate that epistasis cannot be ruled out as an important factor.10 Particularly, results from the UK Biobank are compatible with an upper bound of epistasis explaining slightly more than half as much as additive variance, and a point estimate of epistasis explaining a quarter of the amount of variance explained by additively acting loci. Further extension of epistasis scans, to testing combinations of variants from disease susceptibility regions against the whole genome, or even to genome-wide scans with different, a-priori defined functional, information-based filters, might discover new epistatic interactions,23 thereby, improving both our understanding of disease aetiology and possibly prediction models.

Supplementary material

Supplementary material is available at Cardiovascular Research online.

Acknowledgements

We are grateful for the authorization to apply the data from UK Biobank (project ID: 25214). Data for HiGen were obtained via the database of Genotypes And Phenotypes (dbgap)27 (project ID #49717-3).
Conflict of interest: none declared.

Authors’ contributions

LZ, N.M.S., B.M.M., H.S., contributed to conception and design, TA., T.A., T.F.M.A., contributed to data interpretation. S. Moser, S. Moebus, A.P., K.S., M.M.-Nurasyid, C.G., T.M., E.S.T., W.M., A.M., L.Z., N.M.S., B.M.M., H.S., contributed to conception and design. T.A., T.F.M.A., B.M.M., H.S., T.A., T.F.M.A., contributed to data interpretation. S. Moser, L.Z., N.M.S., J.N.S., B.M.M., H.S., drafted the manuscript. All authors participated in revising it critically for important intellectual content.

Data availability

The data from the German Myocardial Infarction Family Studies (GerMIFS) I,17 II,18 III (KORA),19 IV,20 V,21 VI22 cannot be shared publicly.

The data from the German Myocardi al Infarction Family Studies (GerMIFS) I,17 II,18 III (KORA),19 IV,20 V,21 VI22 cannot be shared publicly.

The data from the German Myocardi al Infarction Family Studies (GerMIFS) I,17 II,18 III (KORA),19 IV,20 V,21 VI22 cannot be shared publicly.

The data from the German Myocardi al Infarction Family Studies (GerMIFS) I,17 II,18 III (KORA),19 IV,20 V,21 VI22 cannot be shared publicly.

The data from the German Myocardi al Infarction Family Studies (GerMIFS) I,17 II,18 III (KORA),19 IV,20 V,21 VI22 cannot be shared publicly.

The data from the German Myocardi al Infarction Family Studies (GerMIFS) I,17 II,18 III (KORA),19 IV,20 V,21 VI22 cannot be shared publicly.

The data from the German Myocardi al Infarction Family Studies (GerMIFS) I,17 II,18 III (KORA),19 IV,20 V,21 VI22 cannot be shared publicly.

The data from the German Myocardi al Infarction Family Studies (GerMIFS) I,17 II,18 III (KORA),19 IV,20 V,21 VI22 cannot be shared publicly.

The data from the German Myocardi al Infarction Family Studies (GerMIFS) I,17 II,18 III (KORA),19 IV,20 V,21 VI22 cannot be shared publicly.

The data from the German Myocardi al Infarction Family Studies (GerMIFS) I,17 II,18 III (KORA),19 IV,20 V,21 VI22 cannot be shared publicly.

The data from the German Myocardi al Infarction Family Studies (GerMIFS) I,17 II,18 III (KORA),19 IV,20 V,21 VI22 cannot be shared publicly.

The data from the German Myocardi al Infarction Family Studies (GerMIFS) I,17 II,18 III (KORA),19 IV,20 V,21 VI22 cannot be shared publicly.

The data from the German Myocardi al Infarction Family Studies (GerMIFS) I,17 II,18 III (KORA),19 IV,20 V,21 VI22 cannot be shared publicly.

The data from the German Myocardi al Infarction Family Studies (GerMIFS) I,17 II,18 III (KORA),19 IV,20 V,21 VI22 cannot be shared publicly.

The data from the German Myocardi al Infarction Family Studies (GerMIFS) I,17 II,18 III (KORA),19 IV,20 V,21 VI22 cannot be shared publicly.
15. Ritchie MD, Steen KV. The search for gene-gene interactions in genome-wide association studies. Nat Rev Genet 2006;7:239–251.

16. Kam-Thong T, Azencott C-A, Cayton L, Pu¨tz B, Altmann A, Karbalai N, Sa¨mann PG, L. Zeng 3. Myers RH, Kiely DK, Cupples LA, Kannel WB. Parental history is an independent risk factor for coronary artery disease. The Framingham Study. Am J Hum Genet 1993;53:648–657.

17. Morgan MD, Pairo-Castineira E, Rawlik K, Canela-Xandri O, Rees J, Sims D, Tenesa J, Sheikhbahaei S, Shepard DS, Shibuya K, Shinohara Y, Shishani K, Shiue I, Shivakoti DL, Tleyjeh IM, Tonelli M, Topouzis F, Towbin JA, Toyoshima H, Traebert J, Tran BX, Trueletal. Tjiuilio L. Tsima Dimbouzea Z, Talibimar B, Tuzcu EM, Ubeda C, Uchidea U, Ukwuona KN, Undurraga EA, Valley AJ, Viver S ja De Gool CH, Van Varakin YY, Vasanak T, Vjzskoppt T, Vlaktea FS, Vanclsa VH, Wagas VH, Walla St, Walsen G, Wrang L, Xiga S, Yang Y, Yano S, Yatsuha Y, Yip P, Yuonemota N, Yoono J, Yungo N, Yu Y, Yun J, Zaki MES, Zalmahshy F, Zeep B, Zhang G, Zhao Y, Zheng Y, Zhu G, Zhi S, Zono S, Zou DX, Zun J, Yasja Y, Zucjo A, Zuhja S, Zheny S, Zhi J.

18. Myers RH, Kiely DK, Cupples LA, Kannel WB. Parental history is an independent risk factor for coronary artery disease: the Framingham Study. Am Heart J 1993;126:157–157.

19. Morgan MD, Pairo-Castineira E, Rawlik K, Canela-Xandri O, Rees J, Sims D, Tenesa J, Sheikhbahaei S, Shepard DS, Shibuya K, Shinohara Y, Shishani K, Shiue I, Shivakoti DL, Tleyjeh IM, Tonelli M, Topouzis F, Towbin JA, Toyoshima H, Traebert J, Tran BX, Trueletal. Tjiuilio L. Tsima Dimbouzea Z, Talibimar B, Tuzcu EM, Ubeda C, Uchidea U, Ukwuona KN, Undurraga EA, Valley AJ, Viver S ja De Gool CH, Van Varakin YY, Vasanak T, Vjzskoppt T, Vlaktea FS, Vanclsa VH, Wagas VH, Walla St, Walsen G, Wrang L, Xiga S, Yang Y, Yano S, Yatsuha Y, Yip P, Yuonemota N, Yoono J, Yungo N, Yu Y, Yun J, Zaki MES, Zalmahshy F, Zeep B, Zhang G, Zhao Y, Zheng Y, Zhu G, Zhi S, Zono S, Zou DX, Zun J, Yasja Y, Zucjo A, Zuhja S, Zheny S, Zhi J.

18. Myers RH, Kiely DK, Cupples LA, Kannel WB. Parental history is an independent risk factor for coronary artery disease: the Framingham Study. Am Heart J 1993;126:157–157.
Translational perspective

Genetic variants identified by GWAS studies explain about a quarter of the heritability of coronary artery disease by additive genetic effects. Our study demonstrates that non-additive effects contribute to the genetic architecture of the disease and identifies complex interaction patterns at the LPA locus, which affect LPA expression, Lp(a) plasma levels, and risk of atherosclerosis. This proof-of-concept study encourages systematic searches for epistatic interactions in further studies to shed new light on the aetiology of the disease.