Clinical and Population Studies

Prediction of Causal Candidate Genes in Coronary Artery Disease Loci


Objective—Genome-wide association studies have to date identified 159 significant and suggestive loci for coronary artery disease (CAD). We now report comprehensive bioinformatics analyses of sequence variation in these loci to predict candidate causal genes.

Approach and Results—All annotated genes in the loci were evaluated with respect to protein-coding single-nucleotide polymorphism and gene expression parameters. The latter included expression quantitative trait loci, tissue specificity, and miRNA binding. High priority candidate genes were further identified based on literature searches and our experimental data. We conclude that the great majority of causal variations affecting CAD risk occur in noncoding regions, with 41% affecting gene expression robustly versus 6% leading to amino acid changes. Many of these genes differed from the traditionally annotated genes, which was usually based on proximity to the lead single-nucleotide polymorphism. Indeed, we obtained evidence that genetic variants at CAD loci affect 98 genes which had not been linked to CAD previously.

Conclusions—Our results substantially revise the list of likely candidates for CAD and suggest that genome-wide association studies efforts in other diseases may benefit from similar bioinformatics analyses. (Arterioscler Thromb Vasc Biol. 2015;35:2207-2217. DOI: 10.1161/ATVBHA.115.306108.)

Key Words: coronary artery disease ■ genome-wide association study ■ microRNAs ■ single-nucleotide polymorphism ■ systems biology

The most recent meta-analysis of genome-wide association studies (GWAS) for coronary artery disease (CAD) identified 46 genome-wide significant and 104 genome-wide suggestive loci associated with increased risk.1 Together these loci explain ∼10% of the heritability. Although quantitatively the effects of common risk alleles identified by GWAS, for example, in the HMGCoR, LDLR or PCSK9 genes, are modest, the pathways tagged by these genes have utmost clinical importance as they constitute prime targets for preventive medication. Accordingly, the largest scientific relevance of GWAS discoveries is seen in elucidation of yet unknown causal mechanisms leading to CAD in the human population.

See accompanying editorial on page 2079

Twelve of the genome-wide significant loci are associated with blood lipid levels and 5 with blood pressure, suggesting that they function through these intermediate phenotypes to increase the risk for CAD.3 However, the

Received on: June 26, 2015; final version accepted on: August 5, 2015.
From the Institut für Integrative und Experimentelle Genomik, Universität zu Lübeck, Lübeck, Germany (I.B., B.R., J.E.); DZHK (German Research Centre for Cardiovascular Research), Partner Site Hamburg/Lübeck/Kiel, Lübeck, Germany (I.B., B.R., J.E.); University Heart Center Lübeck, Lübeck, Germany (I.B., B.R., J.E.); Departments of Medicine (M.C., A.J.L.) and Integrative Biology and Physiology (Y.Z., X.Y.), University of California, Los Angeles; Deutsches Herzcenter München, Klinik für Herz-und Kreislauferkrankungen, Technische Universität München, Munich, Germany (B.V., L.Z., H.S.); Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY (A.D.N., K.H., E.E.S., J.L.M.B.); Icahn Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY (A.D.N., K.H., E.E.S., J.L.M.B.); Cardiovascular Epidemiology and Human Genomics Branch, National Heart, Lung, and Blood Institute, The Framingham Heart Study, Framingham, MA (A.D.J.); Unité Mixte de Recherche en Santé (UMR_S) 1166, Institut National pour la Santé et la Recherche Médicale (INSERM), Paris, France (V.C., D.-A.T.); UMR_S 1166, Team Genomics and Pathophysiology of Cardiovascular Diseases, Sorbonne Universités, Université Pierre et Marie Curie (UPMC Univ Paris 06), Paris, France (V.C., D.-A.T.); Institute for Cardiometabolism and Nutrition (ICAN), Paris, France (V.C., D.-A.T.); Department of Cardiovascular Sciences, University of Leicester, and NIHR Leicester Cardiovascular Biomedical Research Unit, BHF Cardiovascular Research Centre, Leicester, United Kingdom (T.R.W., S.E.H., N.J.S.); Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden (H.F.A., J.L.M.B.); and Department of Molecular and Experimental Medicine, Scripps Research Institute, La Jolla, CA (E.E.S.).

* I. Brænne, M. Civelek, and B. Vilne are joint first authors.
† H. Schunkert and A.J. Lusis are joint senior authors.
‡ A list of CAD Genomics investigators are given in the online-only Data Supplement.

The online-only Data Supplement is available with this article at http://atvb.ahajournals.org/lookup/suppl/doi:10.1161/ATVBHA.115.306108/-/DC1.
Correspondence to Aldons J. Lusis, PhD, 650 Charles E. Young Dr S, Department of Medicine, Division of Cardiology, University of California, Los Angeles, Los Angeles, CA 90095. E-mail jlusis@mednet.ucla.edu
© 2015 American Heart Association, Inc.
Arterioscler Thromb Vasc Biol is available at http://atvb.ahajournals.org DOI: 10.1161/ATVBHA.115.306108

2207
precise genetic mechanisms at the CAD loci affecting either intermediary traits or as of today unknown pathways leading to disease are largely unknown. Thus, translating GWAS loci into genes and pathways will help to provide novel insight into disease susceptibility and ultimately lead to novel treatments for patients with CAD that may also be tailored to the genetic and molecular makeup of individual patients.2

In theory, the process of moving from an associated genetic variant to a disease mechanism seems straightforward and linear: first, identify the causal (rather than the associated) variant, next identify how the causal variant alters gene function of the putative causal gene, and then work out how the altered function of an affected gene perturbs processes at the molecular, cellular, physiological, and whole organism levels that ultimately promote the development of atherosclerosis. However, barring a few notable examples where relevant intermediary phenotypes (eg, an effect on plasma cholesterol) already pointed to a pathway,3,4 experience has shown that dissecting the mechanism is complex. Most lead single-nucleotide polymorphisms (SNPs) identified by GWAS map outside protein-coding regions. Rather, accumulation of lead SNPs was found in regulatory elements that have been identified, for example, in the Encyclopedia of DNA Elements (ENCODE) project.5–7 These results suggest that common genetic variants causally involved in CAD alter gene expression rather than protein sequence.

A further challenge lies in the identification of the causal gene out of multiple candidates in the vicinity of a lead SNP. For example, gene targeting studies in a rat model of hypertension showed that 5 of the 6 genes at the Agrp2-Plod1 hypertension locus affected blood pressure.8 Recent evidence from the Genotype-Tissue Expression project also showed that numerous GWAS loci are associated with the expression of a gene that is not the physically most proximal gene to the locus.9 Even if a single causal gene is identified, defining the function of the gene in the disease context if the gene is poorly annotated is also difficult.

A systematic attempt predicting candidate causal genes and their functional mechanisms in all 159 CAD genome-wide significant and suggestive loci has not yet been carried out. In this study, we used a bioinformatics approach (Figure 1) to (1) catalog all the transcript-coding genes in these loci, (2) evaluate structural SNPs in the protein-coding genes, (3) identify expression quantitative trait loci (eQTL) that overlap CAD loci, and (4) prioritize candidate genes with respect to their likely functional relevance based on evidence from the literature and experimental results from our previous studies.

Materials and Methods

Materials and Methods are available in the online-only Data Supplement.

Results

Locus Boundaries and Genes Within Loci

As a first step, we defined the boundaries of each of the 159 CAD loci by determining the locations of proxy SNPs ($r^2>0.8$) on either side of the lead SNP (Figure 1A in the online-only Data Supplement). There were 3432 proxy SNPs based on the 1000 genomes EUR genotypes determined by the SNAP bioinformatics tool.11 The SNPs were located in 135 nonoverlapping regions. Nine lead SNPs loci had no proxy SNPs suggesting that they probably represent the causal SNP (Table I in the online-only Data Supplement). The size of the remaining regions ranged from 488 bp to 566 kb with an average of 76.5 kb.

To catalog the genes within these CAD loci, we searched the latest release of the ENSEMBL database (release 75), which contains gene model annotations from RefSeq,12 GENCODE,13 and ENSEMBL/HAVANA14 databases. Collectively, within the boundaries of the CAD loci, there were 183 protein-coding genes along with genes for 29 antisense transcripts, 27 long intervening noncoding RNAs, 15 miscellaneous RNAs, 12 miRNAs, 8 small nuclear RNAs, 8 small nuclear RNAs, 3 intronic sense RNAs, 3 processed transcripts that do not contain an open reading frame, 2 long noncoding transcripts that contain a coding gene in their introns on the same strand, and 1 ribosomal RNA (Figure IB in the online-only Data Supplement).

SNP Prioritization Pipeline Overview

To identify the most plausible causal SNP for each locus, we developed an SNP prioritization pipeline (Figure 1). The pipeline consists of 2 main parts: identification of candidate SNPs (Figure 1A) and the identification of genes functionally related to those SNPs (Figure 1B). We considered the 159 peak GWAS SNPs and the 3432 SNPs in high linkage disequilibrium (LD) with the peak SNPs. We assessed the functional implication of each of these SNPs based on 3 main criteria: (1) we checked whether the SNP cause a deleterious amino acid (AA) change; (2) we identified all SNPs that have an eQTL effect, and (3) we identified SNPs that lie within known regulatory regions of the genome. We further analyzed the SNPs that fulfilled at least one of these criteria as potential causal SNPs. The results of the annotation are shown in Table II in the online-only Data Supplement.

Each potential causal SNP assembled by these criteria and the gene it affects either because of AA change or eQTL were next analyzed in the second part of the pipeline. To establish a link between the disease and the candidate genes, we assessed all relevant information from published resources as well as experimental evidence from our laboratories.

CAD Loci With Predicted Nonsynonymous/Deleterious Mutations

To determine SNPs that have protein-altering effects, we annotated all SNPs using ANNOVAR software.15 The gene annotation was based on RefGene, KnownGene, and several functional prediction scores such as SIFT, PolyPhen, and Mutation Taster using the annovar LJB2 database16 as well as the CADD database.17
Bræne et al. Prediction of Causal Genes at CAD GWAS Loci

Of the 159 lead SNPs, 33 (or 20.7%) are exonic: 11 synonymous and 22 nonsynonymous (either the lead SNP itself or its proxy SNPs; Table III in the online-only Data Supplement); 66.7% of all exonic SNPs lead to nonsynonymous AA changes in CAD loci. This percentage was not statistically significantly enriched compared with all known nonsynonymous common SNPs (>1% minor allele frequency) in exonic regions as identified in the 1000 genomes data (48.9% of all exonic SNPs, Fisher $P=0.053$). Next, we focused on the deleterious SNPs because they are likely to be causal. Five lead SNPs (rs3184504, rs3825807, rs867186, rs2571445, and rs11556924) cause AA changes. Three of these are predicted to be deleterious. In addition, 2 of these had proxy SNPs that also cause deleterious AA changes. Furthermore, for 7 other lead SNPs, we identified proxy SNPs that are also predicted to be deleterious. In total, 12 SNPs predicted to be deleterious represent 10 independent loci.

Our results demonstrate the complexity for some of the loci. For example, 2 SNPs (rs1137524 and rs1060407) in the chromosome 3p21.31 locus affect the gene MAP4 complicating the identification of the causal SNP. Furthermore, the lead SNP, rs867186, in the 20q11.22 locus causes a deleterious AA change in the PROCR gene. This SNP is also in high LD with rs11906160 ($r^2=0.92$), which causes a deleterious AA change in the MYH7B gene making it difficult to identify if one or both of these genes are causal for CAD in this locus.

CAD Loci With Regulatory Effects on Gene Expression

To determine SNPs that have effects on gene expression, we interrogated several eQTL results that are part of the Genome-Wide Repository of Associations between SNPs and Phenotypes (GRASP) database.

Figure 1. Candidate single-nucleotide polymorphism (SNP) and gene prioritization pipeline. A, SNPs in 159 coronary artery disease (CAD) genome-wide association study (GWAS) loci were interrogated for their effects on amino acid sequence, gene expression and possible effects on transcription factor binding because of their presence in regulatory regions identified in Encyclopedia of DNA Elements (ENCODE) and National Institutes of Health Roadmap Epigenome projects. B, Genes that were functionally linked to GWAS loci were prioritized based on previous knowledge- or data-driven approaches. DEPICT indicates Data-driven Expression Prioritized Integration for Complex Traits; eQTL, expression quantitative traitloci; GRAIL, Gene Relationships Among Implicated Loci; and MGD, Mouse Genome Database.
Gene Expression (STAGE) study, Massachusetts General Hospital liver/adipose study, and Cardiogenics consortium monocyte/macrophages study, and aortic endothelial cells study. These eQTL results are from >50 tissues and cell types, some are highly relevant to atherosclerosis, such as liver, adipose, and vessel wall as well as monocytes, macrophages, and endothelial cells. We looked for significant association between CAD loci and nearby gene expression (within 1 MB of the lead SNP). We eliminated the spurious associations by assessing if the most significantly associated expression SNP is among the 3591 CAD SNPs. The significant associations and the source of the eQTL association are presented in detail in Table II in the online-only Data Supplement. In total, we found significant associations between 66 CAD lead SNPs and nearby gene expression. This is in contrast to 10 CAD loci found significant associations between 66 CAD lead SNPs and nearby gene expression (within 1 MB of the lead SNP). We eliminated the spurious associations by assessing if the most significantly associated expression SNP is among the 3591 CAD SNPs. The significant associations and the source of the eQTL association are presented in detail in Table II in the online-only Data Supplement. In total, we found significant associations between 66 CAD lead SNPs and nearby gene expression. This is in contrast to 10 CAD loci found significant associations between 66 CAD lead SNPs and nearby gene expression. This is in contrast to 10 CAD loci that are predicted to harbor common coding variants predicted to be deleterious, suggesting that the mechanism of majority of the CAD loci is by regulating nearby gene expression. Another example of complexity was the presence of multiple eQTLs in CAD loci. For example, CAD risk SNP rs17514846 in the 5q26.1 locus is located in the intron of the FURIN gene and is associated with its expression but also with 2 nearby genes, FES and MAN2A2.

### Multiple eQTL Genes in CAD Loci

Another example of complexity was the presence of multiple eQTL genes in CAD loci. In more than half of the loci (38 of 66), the risk SNP affected expression of multiple genes suggesting that several mechanisms, perhaps functioning in different tissues, could be influencing the disease susceptibility. For example, CAD risk SNP rs12936587 in the chromosome 17p11.2 locus lies between genes RAII, PEMT, and RASD1 but is associated with the expression of TOM1L2 located ≈200 kb away from the SNP. Also the CAD risk SNP rs9608859 (22q12.2) that has been traditionally linked to its closest genes OSM or GATS3 is associated with the expression of another gene, SF3A1, ≈65 kb away.

In addition to SNPs located in nongenic regions, we also found SNPs located within a gene but associated with the expression of another nearby gene. For example, rs2681472 (12q21.33) is located in the intron of ATP2B1 but is associated with the expression of GALNT4 ≈1kb away. Another example is the CAD SNP rs2895811 (14q32.2), located in the intron of HHIP1, which is associated with the expression of the YY1 gene in CD19 B cells that lies ≈500 kb away. The region spans more than a dozen genes and miRNAs but the only identified potential disturbed mechanism of this locus is the expression level of YY1.

### Expression SNPs Located in Promoters

SNPs located in promoter or enhancer regions are likely causal variants for regulating the gene expression levels. Sixty-one CAD SNPs with eQTL effects are located in promoter histone marks based on HaploReg annotation (Table II in the online-only Data Supplement). For example, the CAD risk SNP rs590121 in the chromosome 11q13.4 locus is associated with the expression of the CAD loci is by regulating nearby gene expression.

### Table. Predicted Deleterious CAD SNPs

<table>
<thead>
<tr>
<th>Proxy SNP</th>
<th>Lead SNP</th>
<th>Chr</th>
<th>SNP Position (hg19)</th>
<th>Gene</th>
<th>Transcript ID</th>
<th>Amino Acid Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs35107735*</td>
<td>rs12125501</td>
<td>1</td>
<td>169390957</td>
<td>Ccdc181</td>
<td>NM_021179</td>
<td>p.F238I</td>
</tr>
<tr>
<td>rs2820312</td>
<td>rs2820315</td>
<td>1</td>
<td>201869257</td>
<td>Lmod1</td>
<td>NM_012134</td>
<td>p.T295M</td>
</tr>
<tr>
<td>rs35212307*</td>
<td>rs2351524</td>
<td>2</td>
<td>203765756</td>
<td>Wdr12</td>
<td>NM_018256</td>
<td>p.I75V</td>
</tr>
<tr>
<td>rs2571445</td>
<td>rs2571446</td>
<td>2</td>
<td>218831354</td>
<td>Tns1</td>
<td>NM_022648</td>
<td>p.W1197R</td>
</tr>
<tr>
<td>rs1137524*</td>
<td>rs7642590</td>
<td>3</td>
<td>47956424</td>
<td>Map4</td>
<td>NM_001134364</td>
<td>p.V628L</td>
</tr>
<tr>
<td>rs1060407</td>
<td>rs7642590</td>
<td>3</td>
<td>47956037</td>
<td>Map4</td>
<td>NM_001134364</td>
<td>p.S427Y</td>
</tr>
<tr>
<td>rs11556924*</td>
<td>rs11556924</td>
<td>7</td>
<td>129663496</td>
<td>Zc3hc1</td>
<td>NM_016478</td>
<td>p.R363H</td>
</tr>
<tr>
<td>rs1152807</td>
<td>rs7074064</td>
<td>10</td>
<td>88353779</td>
<td>Bmpr1a</td>
<td>NM_004329</td>
<td>p.P2T</td>
</tr>
<tr>
<td>rs1169288*</td>
<td>rs2246068</td>
<td>12</td>
<td>121416650</td>
<td>Hnf1a</td>
<td>NM_000545</td>
<td>p.I27L</td>
</tr>
<tr>
<td>rs4584886</td>
<td>rs4299203</td>
<td>17</td>
<td>17896205</td>
<td>Lrrc48</td>
<td>NM_00130092</td>
<td>p.R191W</td>
</tr>
<tr>
<td>rs11906160*</td>
<td>rs867186</td>
<td>20</td>
<td>33565755</td>
<td>Mvyh7b</td>
<td>NM_020884</td>
<td>p.A25T</td>
</tr>
<tr>
<td>rs867186</td>
<td>rs867186</td>
<td>20</td>
<td>33764554</td>
<td>Prkrc</td>
<td>NM_006404</td>
<td>p.S219G</td>
</tr>
</tbody>
</table>

Of the 29 CAD-related nonsynonymous SNPs, 12 are predicted to have a functional effect. CAD indicates coronary artery disease; and SNP, single-nucleotide polymorphism.

*SNPs indicate that the lead SNP is the potential functional SNP.
with the expression of SERPINH1 and lies in the promoter of the same gene suggesting that the risk SNP alters the binding of transcription factors affecting SERPINH1 gene expression levels and is likely the causal SNP. However, rs2028900, in high LD ($r^2=0.93$) with the CAD risk SNP rs1561198 (2p11.2 locus), lies in the promoter region of MAT2A and is associated with its expression but it is also associated with the expression of nearby genes VAMP8, VAMP5, and GGCX, making it difficult to predict the causal relationship between the risk SNP and nearby genes.

**Tissue-Specific eQTL Effects**

In 2 loci, we observed tissue-specific effects of the risk SNPs. SNP rs602633, located between PSRC1 and CELSR2 genes, in the 1p13 locus has been associated with CAD and lipid levels.25 This locus regulates the expression level of PSRC1, CELSR2, and SORT1 genes in the human liver. In a recent study, hepatic expression of SORT1 has been shown to regulate lipid levels and, therefore, this gene has been predicted to be the causal gene in this locus.3 We observed that the lipid-lowering T allele (of rs602633) is associated with the higher expression of CELSR2 in liver tissue but is associated with lower expression levels in adipose tissue (Figure 2A) suggesting that tissue-specific eQTL effects need to be considered when dissecting the mechanisms of GWAS loci.

Similarly, in the chromosome 17p11 locus, rs4299203 has a suggestive association with CAD (Figure 2B). Expression levels of 5 genes, DRG2, C17orf39, MYO15A, TOM1L2, and SREBF1, are associated with this locus in various tissues (Table II in the online-only Data Supplement). The CAD risk allele (G) is associated with higher level SREBF1 expression in monocytes but is associated with lower expression in macrophages (Figure 2B). SREBF1 encodes the sterol regulatory element-binding protein 1 (SREBP-1), 1 of the 2 major transcription factors that regulate cellular cholesterol levels. This locus is not associated with plasma lipid levels25 suggesting that the SNP effect on SREBF1 is independent of the possible effects of the locus on lipid levels.

**CAD SNPs Affecting miRNA Binding**

A possible mechanism by which risk SNPs affect gene expression is altering the affinity of miRNA binding to the 3′ untranslated regions (UTRs) of disease genes. For example, rs12190287 in the 6q23.2 locus resides in the 3′ UTR of the TCF21 gene and affects binding of miR-224.26 Therefore, we interrogated the CAD SNPs mapping to 3′ UTR region of genes using the microSNIPER database to assess whether they could reside in a predicted target miRNA-binding site.27 We restrained our analysis to miRNAs with predicted seed length of 7mers or more. Fifty-five 3′ UTR CAD SNPs from 33 distinct genes were predicted to lie within an miRNA-binding site for a total of 254 distinct miRNAs (Table IV in the online-only Data Supplement). The predicted number of miRNAs targeting the 3′ UTR region of a gene ranged from 1 (for BCAP29, MAP4, RND3, and WDR12) to 29 (for MRAS). Of note, 23 miRNAs were predicted to bind >1 candidate
CAD gene. For example, hsa-miR-130a-5p was predicted to bind \textit{UBE2Z} (with the 3′ UTR SNP rs15563) and \textit{SLC22A3} (with the 3′ UTR SNP rs3088442), and hsa-miR-4722-5p was predicted to bind \textit{APOA5} (with the 3′ UTR SNP rs2266788) and \textit{ICA1L} (with the 3′ UTR SNP rs72932707). In accordance with the expected effect on \textit{APOA5}, rs2266788 was significantly associated with plasma triglyceride levels.\textsuperscript{25}

Of the 55 SNPs affecting miRNA binding, 13 are also associated with the expression of the same gene. At the 11p15.4 locus, rs360137 affects the binding of hsa-miR-3198 to the 3′ UTR of the \textit{SWAP70} gene and is also associated with the expression of the same gene. Similarly, rs1058588 at the 2p11.2 locus affects the binding of hsa-miR-5197-3p to the 3′ UTR of \textit{VAMP8} and is also associated with the \textit{VAMP8} expression. SNP rs12733378 at the 1q32.1 locus affects the binding of 5 miRNAs at the 3′ UTR of \textit{CAM5AP2} and is also associated with the expression of the same gene. These examples suggest that eQTL effect may be because of altering miRNA binding to the target genes.

**SNPs Affecting miRNA Binding and Promoter Regions**

We also observed that SNPs that altered miRNA-binding sites in \textit{UBE2Z} and \textit{MAP4} were in high LD with SNPs in their promoter regions. rs6442101 is predicted to be in the promoter region of \textit{MAP4} in various cell types and tissues examined in the ENCODE and NIH RoadMap Epigenome project.\textsuperscript{10} This SNP is in high LD with rs1061003 ($r^2=0.97$), which is predicted to affect the binding of miR-378a-5p in the 3′ UTR of \textit{MAP4}. Similarly, rs999474 located in the promoter region of \textit{UBE2Z} is in high LD with rs15563 ($r^2=0.84$), which is predicted to affect the binding of 8 different microRNAs (Table IV in the online-only Data Supplement). Therefore, it is plausible that the \textit{MAP4} and \textit{UBE2Z} loci affect the expression of these genes by either altering the affinity of transcription factor binding in the promoter region or miRNA binding in the 3′ UTR region. By studying the expression patterns of transcription factors or miRNAs whose binding would be altered, it might be possible to predict the tissue which these genes would be functional in the context of CAD.

**Candidate Gene Prioritization and Prediction of Novel CAD Genes**

The CAD GWAS loci have been typically annotated based on their proximity to a gene, yielding a total of 161 genes. However, recent literature evidence\textsuperscript{26,28,29} suggests that the nearest gene is often not the target of a given GWAS association. Instead, the identification of eQTLs can be used for predicting the target genes. In this work, annotating the 159 CAD risk SNPs led to a list of 151 CAD candidate risk genes based on nonsynonymous AA changes and eQTL effects. Of note, we were not able to assign a gene to all loci. We compared our list of genes with the GWAS genes reported in the literature\textsuperscript{1} and identified 98 CAD genes hitherto not considered to be involved in the genetics of CAD for which our bioinformatics data provide suggestive evidence (Table V in the online-only Data Supplement; Figure 3A). Of the previously considered GWAS genes, 108 do not overlap with our annotation. These genes might be unrelated to CAD or missed by our annotation efforts. We attempted to prioritize the 98 novel genes using literature and database-based approaches (previous knowledge) or using analyses performed in this article and data from our laboratories (data driven).

For the previous knowledge approach, we first used a statistical text mining approach, Gene Relationships Among Implicated Loci (GRAIL),\textsuperscript{30} that assesses the degree of relatedness among putative candidate genes within disease regions using PubMed article abstracts. Second, we used another integrative tool, Data-driven Expression Prioritized Integration for Complex Traits (DEPICT),\textsuperscript{31} that predicts gene functions from manually curated pathways, protein–protein interaction screens and phenotypes from mouse gene knockout studies to prioritize the most likely causal genes at each associated locus, as well as performs pathways enrichment analysis and identifies tissues and cell types where genes from the associated loci are highly expressed. Third, we used the functional annotation information available from the public databases: (1) Mouse Phenotypes from the Mouse Genome Database,\textsuperscript{32} (2) Functional Disease Ontology,\textsuperscript{33} and (3) Biochemical Pathway information, as collected from the ConsensusPathDB database\textsuperscript{34} and the gene ontology annotation.\textsuperscript{35,36} If a gene was predicted to be a causal gene (eg, \textit{Ps}$\leq0.05$ assigned by GRAIL or DEPICT) or its functional annotation (eg, Mouse Phenotype, Biochemical Pathways, or Disease/gene ontology annotation) was CAD related (Methods and Data for the definition of CAD relatedness section in the online-only Data Supplement), we assigned a score of 1 for a total of 6 (Table VI in the online-only Data Supplement). Of note, the previous knowledge-driven approaches are biased because of the availability of literature-based information on well-studied genes. As a result, using previous knowledge-driven approach, 4 genes reached the maximum score of 6: \textit{LPL}, \textit{CDKN2B}, \textit{ALDH2}, and \textit{PROC}, all of them being among the 604 genes with CAD-related evidence manually extracted from scientific publications and deposited in the Coronary Artery Disease Gene Database (CADgene) V2.0.\textsuperscript{17}

Because of the biased previous knowledge scoring, we also used an alternative, data-driven, approach to look for novel candidate genes. For the data-driven approach, we assigned scores to the genes if they harbored nonsynonymous SNPs, had eQTLs, had promoters with CAD SNPs, were members of a CAD-relevant Bayesian Network constructed from CAD-relevant tissue gene expression studies or had a significant correlation with aortic root lesion size in a systems genetics study of atherosclerosis in a mouse population.\textsuperscript{37} Hence, the total score a gene could achieve was 5 based on the data-driven approach. We present these prioritization results in Table VI in the online-only Data Supplement. For 69% of the genes, there was evidence from both the previous knowledge-based or data-driven approaches. However, for 31% of the genes, only the data-driven approach provided evidence for the involvement in CAD pathogenesis. The results of the prioritization approach can be found in the Table VI in the online-only Data Supplement. Here, we highlight some of the new potential CAD genes that were prioritized based on our data-driven approach. \textit{REST}, \textit{GIP}, and \textit{TMEM116} received the top 3 scores.
SNP rs17087335 at the 4q12 locus is located on the NOA1 gene but has proxy SNPs that lead to nonsynonymous AA changes in the REST gene. This CAD locus is also associated with the expression of REST in lung tissue and the CAD SNPs are located in its promoter. Furthermore, the aortic expression of REST is significantly correlated with lesion size in mice.37 REST encodes for a transcriptional repressor that has been shown to play a role in the phenotypic modulation of vascular smooth muscle cells.38 REST binds to the promoter of the potassium channel KCa3.1 and represses its expression during intimal hyperplasia. In humans, there seems to be an inverse correlation between REST expression and vascular smooth muscle cell proliferation.34 Consistently, in our mouse data set,37 we observed an inverse correlation between aortic expression of Rest and lesion size (r = −0.24; P = 0.03).

SNP rs15563 at the 17q21.32 locus is located on the UBE2Z gene but has a proxy SNP that lead to nonsynonymous AA changes in the GIP gene. The locus is also associated with the expression of this gene and CALCOCO2, DLX4, SPAG9, ATP5G1, D6RB11, and UBE2Z. GIP is the highest ranked gene for this locus based on our scoring. It is an incretin hormone that belongs to the glucagon superfamily and is associated with insulin secretion.39 Mouse knockout models of Gip showed reduced obesity and insulin resistance.40,41 In our mouse data set,37 we also found positive correlation between Gip expression in the aorta and lesion size (r = 0.23; P = 0.05).

SNP rs3809274 at the 12q24.13 locus has previously been annotated with the ATXN2 gene.1 It is also in high LD with an SNP increasing the expression of the transmembrane protein 116 (TMEM116). Although little is known about this gene, in our data-driven approach, it is one of the highest ranked genes. We do not observe an eQTL effect on ATXN2 but we observe an eQTL effect on TMEM116, C12orf30, SH2B3, BRAP, ALDH2, MAPKAPK5-AS1, HECTD4, and MAPKAPK5. We prioritized TMEM116 based on our data-driven approach because it harbored SNPs with nonsynonymous AA change and its expression level was associated with the CAD locus in multiple tissues (Table VI in the online-only Data Supplement).

In addition, we highlight MYH7B here. The gene is linked to rs867186 at the 22q11.2 by a SNP predicted to cause a deleterious AA change. The lead SNP lies in the PROCR gene and also associates with the expression of this gene. It is also associated with the expression of 8 other genes, including the MYH7B gene in his locus; however, rs867186 is located in the promoter of MYH1B. MYH7B encodes the heavy chain of myosin II and is expressed in heart and skeletal muscle.42 It is also reported to be expressed in smooth muscle cells in mice.43 Therefore, we prioritized MYH7B as the causal candidate gene out of the 8 genes in this locus.

Finally, to assess the information gain of our annotation effort, we compared the genes previously assigned to the loci1 and our annotations (Figure 3B; Table V in the online-only Data Supplement).
Data Supplement). Of the 159 CAD GWAS loci, 15 loci (9%) showed identical annotation. For 25 loci (16%), we found no overlap between our new and the traditional annotation. We could not identify any functional evidence for the genes traditionally assigned to the loci, but we could link other genes with functional evidence. For 32 loci (20%), we gained additional knowledge using our annotation. Of these, 8 loci (5%) are traditionally annotated with at least 1 gene that cannot be validated by our annotation and hence might be wrong. For the other 24 loci (15%), we annotate additional genes that are not reported for the loci to date. For 87 loci, we did not find any functional links and hence no gene assignment.

Discussion
Recent genome-wide association efforts to understand the genetic architecture of coronary atherosclerosis and myocardial infarction have led to the identification of numerous novel DNA variants associated with disease risk. The main challenge for gaining biological insights from genetic associations is identifying which genes and pathways explain the associations. Only a few studies have partially identified the susceptibility mechanisms affected by CAD loci and thereby offer blueprints for subsequent efforts to explain disease pathogenesis. These include CAD risk alleles at the 1p13, 6q23, and 4q32 loci, which displayed functional links to gene expression and related disease mechanisms involving SORT1, TCF21, and GUCY1A3.3,26,44,45

The most robustly associated chromosomal region, the 9p21 locus, still remains a mystery after almost a decade of studies.46 More comprehensive efforts are needed to translate the GWAS loci into actionable genes and pathways. Cell type–specific eQTL or coding (nonsynonymous) variants in strong LD with associated variants can potentially link these variants to genes involved in atherosclerosis. Here, we queried publicly available databases and our own experimental data sets to predict the functional genes in the CAD genome-wide significant and suggestive loci.

We observed that the majority of the GWAS loci affect gene expression as opposed to leading to AA changes (41% versus 6%). This is in agreement with previous studies that predicted 70% to 80% of GWAS SNPs to be regulatory.5,47 Among the loci that lead to changes in gene expression, we revealed that several are associated with differential expression of multiple genes in multiple tissues. Moreover, we observed at a few loci both protein alterations and eQTL effects for significantly associated SNPs. The finding that 1 locus might harbor several proxy SNPs that have eQTL effects on different genes adds further to the complexity of elucidating genetic mechanisms underlying CAD. This is, for example, true for the 19q13.32 locus where the lead SNP is in LD with a missense SNP (in the APOC4 gene) and in LD with SNPs affecting the expression of 3 genes (APOC4, APOC2, and APOE). Second, we identified multiple SNPs that alter promoter and enhancer sequences. ENCODE data indicate that the average number of target genes of a distal regulatory element is 2.5, suggesting that the expression of >1 gene is affected.29 A potential example of this mechanism is the 10q22.3 locus. The CAD SNP lies in a potential enhancer and is associated with the expression of 2 genes, ANXA11 and MATJA. Third, we observed loci that affect the expression of a transcription factor leading to changes in the expression of nearby genes. An example is the 17p11 locus that regulates the expression of the transcription factor SREBP1, as well as 4 other nearby genes. Hence, our annotation efforts show that the downstream effects of a locus may be highly complex, not fitting into 1 pathway or function (eg, the 9p21 locus), and that some loci may contain multiple causal genes. Further efforts to analyze pathways and gene networks affected by individual loci will be useful to understand if >1 gene is functional in a locus.

We relied heavily on eQTL data to identify likely causal genes at the CAD loci. Because a large fraction of the variation underlying common diseases seems to be regulatory,5,6,47 this is a sound strategy. But we note that the sample sizes in eQTL studies vary considerably; hence, there may have been insufficient power to detect the eQTLs at the CAD loci. In addition, there may have been confounding factors contributing to the detection of an eQTL, such as population structure or experimental heterogeneity. Importantly, only few studies used tissues relevant to atherosclerosis, such as endothelial cells22 or the vascular wall,20 to detect eQTLs as these tissue resources are difficult to obtain. Ongoing projects such as the Genotype-Tissue Expression,9 will contribute to the identification of eQTLs in CAD-relevant tissues.

The complex expression patterns of multiple genes regulated by significantly associated SNPs at a single locus make it challenging to dissect the principle mechanism of the locus altering disease risk. A SNP affecting several genes (either the expression or the protein sequence) might increase the risk of the disease in an additive fashion. However, the disease might also only be caused by only 1 of the altered functions. Hence, it is not straightforward to identify the underlying disease mechanisms. In silico prediction can help to establish a link between an identified genetic effect and the disease. However, functional characterization using molecular biology and genetic approaches are required to understand the mechanisms in more detail.

Traditionally for locus annotations, the nearest gene to the identified risk SNP is reported. However, recent evidence suggests that because of the three-dimensional chromosomal confirmation, genomic locations that seem to physically distant can interact with each other.47 One such example is the FTO obesity locus that was shown to interact, at megabase distance, with the enhancer of transcription factor IRX3 and regulate its expression.48 The majority of GWAS-identified variants fall in noncoding regions of the genome, the most frequent elements affected being transcriptional enhancers and silencers, which are typically located >1 kb from their target genes and regulate transcription through long-range interactions.28 In fact, recent analysis by the ENCODE Consortium demonstrated that only ≈27% of the distal regulatory elements have an interaction with the nearest promoter,29 suggesting that the nearest gene is often not the target of a given GWAS association. Therefore, we used local eQTL results from various resources and protein-altering information, and identified 98 genes that had not been linked to the CAD loci previously. However, this
analysis typically provides indirect evidence of an association, and the overlap of an eQTL with a disease locus may be coincidental. Our annotation is also limited because of lack of results from 5C or other chromatin capture methodology to assess long-range genomic interactions in CAD-relevant tissues. It is crucial to consider disease-relevant tissue types as some eQTLs are tissue dependent and trait-associated variants tend to exert more tissue-specific effects. Additional functional assays would be required for confirming the mechanistic relevance of these eQTLs to the disease or trait.

Gene assignment without functional evidence demonstrates the misleading potential of GWAS reports. It is biased by the biological relevance (and reported phenotypic effects) of the neighboring genes. For instance, the locus on chromosome 19p13.2 with the lead SNP rs1122608 spanning the SMARCA4 gene is also assigned to the neighboring LDLR gene, which has a well-established role in regulating plasma low-density lipoprotein (LDL) levels. In this work, we only identify an eQTL link between the locus and the SMARCA4 but not the LDLR gene. Our results imply that the disease-causing effect underlying the locus could be the altered expression of the SMARCA4 gene rather than influencing the LDLR. Alternatively, the tissue samples evaluated here for studying eQTL effects missed the interaction of the SNP and LDLR expression, which is less likely because LDLR is primarily expressed in the liver and our eQTL databases included ample liver eQTLs from multiple studies. After all, the LDLR gene is clearly a causal CAD gene, just perhaps not underlying this GWAS signal. Another example is the lead SNP at locus 12q24.12, rs3809274 (Figure 4). It is located between the genes ATXN2 (upstream) and BRAP (downstream) and was traditionally assigned to ATXN2. However, based on our annotation effort, we do not find a link between the locus and ATXN2, but instead on 6 other genes (Table II in the online-only Data Supplement). Another lead SNP, rs3184504, downstream of ATXN2 and located within SH2B3 is, however, associated with the expression of ATXN2 gene. Of note, rs3184504 and rs3809274 are not in LD. rs3184504 was assigned to SH2B3, but we do not find a functional link between rs3184504 and SH2B3. There is, however, an eQTL association between rs3809274 and SH2B3. In other words, although ATXN2 and SH2B3 are both CAD GWAS candidate genes, the previous gene locus assignment is not supported by our data.

We used previous knowledge and data-driven approaches to prioritize the novel candidate genes. However, we note that using previous biological knowledge about the candidate genes undermines the agnostic nature of the GWAS approach. In addition, both the number of functional annotations per gene and the number of genes per functional annotation demonstrate scale-free properties, meaning that there is a small number of genes with a large number of functional annotations, whereas for a large number of genes there are only few or no functional annotations available. Hence, these approaches are limited by incomplete information about gene functions. From our pipeline, most functional annotations could be retrieved when searching for Biochemical Pathways in the ConsensusPathDB database and for the gene ontology annotations, where 75 of 154 or ≈49% genes could be mapped to at least 1 annotation term. However, this also indicates that about half (≈51%) of the candidate genes lacked any functional annotation and, therefore, could not be considered for prioritization here. Through our prioritization pipeline, we would again select only well-annotated genes for further studies (the rich get richer principle), whereas the biological function of underinvestigated and underannotated genes would further remain enigmatic; therefore, we highlighted some of the novel genes with top scores from only the data-driven prioritization approach.

Finally, we note that, although we were as comprehensive as possible in annotating the CAD loci, we are limited by the available data sets from previous studies. CAD loci may harbor coding variants that are not presently in databases or regulatory variants that may affect gene expression in a tissue or cell type that has not been examined. For example, the CAD loci, as defined by SNPs in high LD with the lead SNP contain a total of 291 genes, ≈40% of which are noncoding. Recent large RNA sequencing studies and integrative projects such as ENCODE suggest that noncoding RNAs constitute ≈60% of transcribed RNAs. Moreover, in recent years functional studies suggest that they play an important role in the regulation of transcription and translation. We do not have microarray probes measuring the expression levels of all protein-coding or noncoding genes. Furthermore, most eQTL studies were of moderate sample size; therefore, the power to detect significant associations is limited. In addition, we only considered SNPs in high LD (r²≥0.8) with the peak SNP. It is possible that the GWAS lead SNP imperfectly tags the causal SNP, which is in moderate LD with the lead SNP, that is, r²<0.8. This could be the case if the causal variant has slightly different minor allele frequency compared with the lead SNP. Then by focusing on SNPs with r²≥0.8 with the lead SNP may lead to improper

![Figure 4. Gene reassignment based on expression quantitative trait loci (eQTL) effects. The lead SNP rs3809274 was traditionally assigned to ATXN2. This link was not verified by our annotation effort. rs3809274 is associated with the expression of SH2B but not of ATXN2. The lead single-nucleotide polymorphism (SNP) rs3184504 was traditionally linked to SH2B, but the functional effect identified in this work, links the SNP to the ATXN2 gene.](image-url)
conclusion about the functional variant and the causal gene. However, if we were to reduce the LD threshold, this could have led to spurious associations even though the eQTL and CAD locus are actually independent from each other but in low LD.

Ultimately, for a full understanding, each CAD locus will have to be individually investigated using tools, such as experimental organisms and iPS cells. In this study, we have used some relatively standard tools to refine the list of candidates. Additional approaches that could be useful at present include chromosome conformation analyses, application of novel algorithms for causal SNP analysis, network analyses, and identification of rare variants. Looking forward, new resources and tools, such as noncoding RNA annotation, RNA-binding maps, splicing variants and code annotation, and detailed enhancer and transcription maps in a variety of cell types relevant to atherosclerosis, will greatly assist such efforts.

Sources of Funding
This work was supported by Transatlantic Networks of Excellence Award (12CVD02) from Foundation Leducq (all authors), US National Institutes of Health grants HL28481, HL114437, HL30568, and HL123295 (A.J. Luisi). Funding is provided by the German Federal Ministry of Education and Research in the context of the e：Med program (e：AtherosysMed and sysINFLAME), the FP7 European Union project and CVgenes@target (261123). Further grants were received by the Universität zu Lübeck. This study was also supported through the Deutsche Forschungsgemeinschaft cluster of excellence Inflammation at Interfaces. M. Civelek is supported by an National Institutes of Health Pathway to Independence Award (2K12HL121172). Y. Zhao and X. Yang are supported by an American Heart Association Scientist Development Grant (15SDG17290032). N.J. Samani and T.R. Webb are funded by the British Heart Foundation. N.J. Samani is a UK National Institute for Health Research (NIHR) Senior Investigator. S.E. Hamby is funded by the NIHR Leicester Cardiovascular Biomedical Research Unit. J.L.M. Rasmussen is the founder, main shareholder, and chairman of the University of Tartu Science Centre-Karolinska Institutet and by the University of Tartu (SPIGVARENG).

Disclosures
Dr Björkergen is the founder, main shareholder, and chairman of the board for Clinical Gene Networks (CGN). Dr Schadt is a board member of CGN.

References
Diabetes alleviates obesity and insulin resistance under high-fat diet conditions. 

Coronary artery disease remains the leading cause of death in the Western world despite significant advances in early detection and extensive use of lipid-lowering and antihypertensive drugs. The pathogenesis of atherosclerosis involves environmental factors, and hundreds of genetic variants and their interactions, each exerting a relatively small effect on disease susceptibility. A more complete understanding of the disease susceptibility is urgently needed to develop additional diagnostics and therapeutics.

Genome-wide association studies identified numerous genetic loci associated with coronary artery disease. Translating these findings into therapies will require the identification of causal genes in the associated loci. In this study, we used publicly available and in-house functional information to systematically review evidence of the involvement of genes in and near the associated loci. Using this approach, we identified 98 possible novel candidate genes to be involved in the pathology of coronary artery disease.

2217

**Significance**

Coronary artery disease remains the leading cause of death in the Western world despite significant advances in early detection and extensive use of lipid-lowering and antihypertensive drugs. The pathogenesis of atherosclerosis involves environmental factors, and hundreds of genetic variants and their interactions, each exerting a relatively small effect on disease susceptibility. A more complete understanding of the disease susceptibility is urgently needed to develop additional diagnostics and therapeutics. Genome-wide association studies identified numerous genetic loci associated with coronary artery disease. Translating these findings into therapies will require the identification of causal genes in the associated loci. In this study, we used publicly available and in-house functional information to systematically review evidence of the involvement of genes in and near the associated loci. Using this approach, we identified 98 possible novel candidate genes to be involved in the pathology of coronary artery disease.
Prediction of Causal Candidate Genes in Coronary Artery Disease Loci
on behalf of the Leducq Consortium CAD Genomics!

Arterioscler Thromb Vasc Biol. 2015;35:2207-2217; originally published online August 20, 2015;
doi: 10.1161/ATVBAHA.115.306108
Arteriosclerosis, Thrombosis, and Vascular Biology is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2015 American Heart Association, Inc. All rights reserved.
Print ISSN: 1079-5642. Online ISSN: 1524-4636

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://atvb.ahajournals.org/content/35/10/2207

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Arteriosclerosis, Thrombosis, and Vascular Biology can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Arteriosclerosis, Thrombosis, and Vascular Biology is online at:
http://atvb.ahajournals.org//subscriptions/
Supplementary Figure I: Definition of CAD locus boundaries (A) and the transcript coding genes in the CAD loci (B)
Supplemental Text

Member of the Leducq CADGenomics Consortium

Redouane Aherrahrou¹,²,³, Zouhair Aherrahrou¹,²,³, Jaafar Al-Hasani¹,²,³, Sarah Andrews⁴, Dimitra Aravani⁴, Charlotte L. Beynon⁴, Johan L.M. Björkegren⁵,⁶,⁷, Yuna Blum⁸, Ingrid Braenné¹,²,³, Peter S. Braund⁴, Francois Cambier⁹,¹⁰,¹¹, Mete Civelek⁸, Veryan Codd⁴, Veronica Codoni⁹,¹⁰,¹¹, Ariella Cohain¹,²,³, Pallavi R. Devchand⁵,⁷, Antonio F. Di Narzo⁵,⁷, Jeanette Erdmann¹,²,³, Hassan Foroughi Asl⁶, Oscar Franzen⁷, Maryam Ghaderi Najafabadi⁴, Ke Wei Gong⁸, Peng Gong⁴, Stephen Hamby⁴, Ke Hao⁵,⁷, Lesca M. Holdt¹², Peter D. Jones⁴, Michael A. Kaiser⁴, Sonia Karabina¹⁰,¹³, Elisavet Karamanavi⁴, Thorsten Kessler¹⁴, Mariana Kleinecke¹,²,³, Aldons J. Lusis¹⁵,¹⁶, David G. McVey⁴, Gavin E. Morris⁴, Evan Muse¹⁷, Christopher P. Nelson⁴, Ewa Ninio⁹,¹⁰,¹¹, Sukhvir K. Rai⁴, Benedikt Reiz¹,²,³, Nilesh J. Samani⁴, Eric E. Schadt⁵,⁷, Heribert Schunkert¹⁴,¹⁸, Erick R. Scott¹⁷, Maria Segura Puimedon¹,²,³, Emma Stringer⁴, Husain Ahammad Talukdar⁵, Daniel Teupser¹², Eric J. Topol¹⁷, Ali Torkamani¹⁷, Zahia Touat⁵,⁷, David-Alexandre Trégouet⁸,¹⁰,¹¹, Zhidong Tu⁵,⁷, Baiba Vilne¹⁴, Simon von Ameln¹⁴, Thomas R. Webb⁴, Xia Yang¹⁹, Shu Ye⁴, Lingyao Zeng¹⁴, Yuqi Zhao¹⁹, Jun Zhu⁵,⁷

¹ Institut für Integrative und Experimentelle Genomik, Universität zu Lübeck, 23562 Lübeck, Germany
² DZHK (German Research Centre for Cardiovascular Research), partner site Hamburg/Lübeck/Kiel, 23562 Lübeck, Germany
³ University Heart Center Luebeck, 23562 Luebeck, Germany
⁴ Department of Cardiovascular Sciences, University of Leicester, and NIHR Leicester Cardiovascular Biomedical Research Unit, BHF Cardiovascular Research Centre, Leicester, UK
⁵ Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA
⁶ Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden
⁷ Icahn Institute for Genomics and Multi-Scale Biology, Icahn School of Medicine at Mount Sinai, New York, New York, USA
⁸ Department of Medicine, University of California, Los Angeles, Los Angeles, 90095, CA, USA
⁹ Institut National de la Santé et la Recherche Médicale (INSERM), Unité Mixte de Recherche en Santé (UMR_S) 1166, Team Genomics & Pathophysiology of Cardiovascular Diseases, F-75013, Paris, France
¹⁰ Sorbonne Universités, Université Pierre et Marie Curie (UPMC Univ Paris 06), Paris, France.
¹¹ Institute for Cardiometabolism and Nutrition (ICAN), F-75013, Paris, France
¹² Institute of Laboratory Medicine, Ludwig-Maximilians-University Munich, Marchioninistr. 15, 81377 Munich, Germany
¹³ INSERM, UMR_S 933, Hôpital Armand-Trousseau, Paris, France
¹⁴ Deutsches Herzzentrum München, Klinik für Herz- und Kreislauffkranckungen, Technische Universität München, Munich, Germany
¹⁵ Department of Microbiology, Immunology, and Molecular Genetics, University of California, Los Angeles, Los Angeles, CA 90095, USA
¹⁶ Department of Human Genetics, University of California, Los Angeles, Los Angeles, CA 90095, USA
¹⁷ Scripps Translational Science Institute, 3344 N. Torrey Pines Court, La Jolla, California 92037, USA
18  DZHK (German Research Centre for Cardiovascular Research), partner site Munich Heart Alliance, Munich, Germany
19  Department of Integrative Biology and Physiology, University of California, Los Angeles, Los Angeles, CA, 90095, USA
Supplemental Material

Materials and Methods
Candidate gene prioritization using prior knowledge from published literature and public databases

GRAIL
Gene Relationships Among Implicated Loci (GRAIL)\(^1\) is a statistical text mining approach that quantifies the relatedness among genes within disease regions, where gene relatedness is defined as the degree of similarity in the text describing them within PubMed article abstracts. PubMed abstracts prior to December 2006 can be selected, in order to avoid publications that report on or are influenced by GWAS papers. In addition to the text-based similarity measure, GRAIL also uses two alternative measures of relatedness, including similarity in Gene Ontology (GO)\(^2\) annotations and gene expression data across multiple human tissues from the Novartis Gene Expression Atlas\(^3\). For each disease region, GRAIL selects the single most connected gene as the key gene and the disease region is assigned that key gene’s p-value. In addition, GRAIL also returns a set of keywords that may provide insight into the underlying biological pathways. Here, a candidate gene received a score of 1, if it had a significant p-value (p≤0.05) in one of the GRAIL’s similarity measure categories in the following order: (1) after the survey of literature prior to December 2006; (2) PubMed abstracts of August 2012; (3) tissue expression or (4) GO annotation. Otherwise, the gene was scored 0.

DEPICT
We used Data-driven Expression Prioritized Integration for Complex Traits (DEPICT)\(^4\), an integrative tool that employs predicted gene functions to prioritize the most likely causal genes at each associated loci, as well as performs pathways enrichment analysis and identifies tissues and cell types where genes from associated loci are highly expressed. DEPICT combines co-regulation of gene expression from 77,840 samples with 14,461 previously annotated gene sets, including manually curated pathways, data from protein-protein interaction screens and phenotypic gene sets from mouse gene knock-out studies, in order to predict gene function based on a ‘guilt-by-association’ procedure. Using these gene sets and a set of trait-associated loci, DEPICT assesses whether any of the gene sets are significantly enriched for genes in the associated loci and prioritizes genes that share predicted functions with genes from the other associated loci more often than expected by chance. In addition, 37,427 human microarrays are being utilized to identify tissue and cell types in which genes from associated loci are highly expressed. Moreover, DEPICT uses precomputed GWAS based on randomly distributed phenotypes (‘null GWAS’) to adjust the p-values from the above three analyses for null expectation. We assigned a candidate gene a score of 1, if DEPICT identified it as a prioritized gene (p ≤ 0.05) in the associated loci, otherwise, the gene was scored 0.

Mouse Phenotype
We downloaded Mouse Phenotype information from the Mouse Genome Database (MGD)\(^5\), which is a comprehensive, non-redundant catalog of mouse genome features obtained through loads from major data providers and experimental consortia, electronic submissions from laboratories and from the biomedical literature. For each of the candidate genes, we first collected all Mouse Phenotypes associated with the gene (Supplementary Table 6). This information was used in two ways. First, we performed Mouse Phenotype enrichment analysis in order to highlight the over-represented categories. Fisher’s exact test, as implemented in R was used to calculate the statistical significance of the overlaps, thereafter the false discovery rate (FDR) was controlled using the Benjamini-Hochberg (BH) procedure\(^6\). Second, Mouse Phenotype annotations were used in order to score the candidate genes within and across the CAD risk loci: each gene received a score of 1, if it was mapped to at least one CAD-relevant Mouse Phenotype, such as ‘MP:0005385-cardiovascular system phenotype’, ‘MP:0005375-adipose tissue phenotype’ or ‘MP:0005387-immune system phenotype’.

Disease Ontology
Functional Disease Ontology (FunDO) is a comprehensive disease to gene annotation, which uses the Disease Ontology (DO) to identify relevant diseases in GeneRIFs. Disease Ontology (DO) annotations of a human gene are complementary to Gene Ontology (GO) Consortium annotations and describe unique roles for genes in the context of disease. A GeneRIF (Gene Reference Into Function) is a brief (up to 255 character) Wiki-type annotation to a gene in the NCBI database and contains gene specific information including disease associations. Again, for each of the candidate genes, we first collected all FunDO disease categories associated with the gene (Supplementary Table 6). Thereafter, these FunDO annotations were used similarly to the Mouse Phenotype information: (1) disease enrichment analysis was performed and (2) each gene received a score of 1, if it was mapped to at least one CAD-relevant disease, e.g., 'Atherosclerosis', 'Cardiovascular disease', 'Hyperglycemia' or 'Hypertension'.

Biochemical Pathways
Biochemical Pathway information was collected from ConsensusPathDB, which is an integrative database currently containing 4,601 Biochemical Pathways in Homo sapiens from 32 public resources such as KEGG, Reactome, Wikipathways and NetPath. In addition, we also included two CAD-related data sets: (1) 604 genes with CAD-related evidence manually extracted from over 5,000 scientific publications from the Coronary Artery Disease Gene Database (CADgene) V2.0, which we designated as 'CADgene'; (2) Cardiovascular Gene Ontology Annotation Initiative uses Gene Ontology (GO) terms to curate scientific literature and integrate results from high-throughput experiments in order to create an informative resource for the cardiovascular-research community. Over 4,000 cardiovascular-associated genes have been identified as targets for annotation with GO terms. We also included these genes in our Pathway set ('Cardiovascular Gene'). Again, we first collected all Pathways associated with each of the candidate genes (Supplementary Table 6) and performed (1) enrichment analysis to highlight the over-represented categories, as described above for the Mouse Phenotype analysis and (2) to score the candidate genes based on their mapping to CAD-relevant pathways. In this case, in order to classify a Pathway as being CAD-relevant, it had to be identified as such by two of the three strategies applied: (1) manual inspection; (2) significantly (FDR ≤ 0.05) over-represented in a list of CAD-related genes from the Coronary Artery Disease Gene Database (CADgene) V2.0; (3) selected as CAD-relevant based on a set of disease-related keywords from the scientific literature (e.g., 'hypercholesterolemia', 'cholesterol pathway', 'vascular disorder'), whereas the keywords were selected by automatically searching the scientific literature using PolySearch, a web-based text mining tool allowing the extraction of relationships between human genes, diseases, pathways, drugs and metabolites. This was followed by manual curation, in order to remove the false positives. In this case, we searched for all entities related to disease ‘Atherosclerosis/CAD/MI’, as well as to two well-known CAD genes apolipoprotein E (APOE) and the low density lipoprotein receptor (LDLR). Finally, this list was used in order to score the candidate genes: each gene received a score of 1, if it was mapped to at least one CAD-relevant Pathway, such as 'Calcium Regulation in the Cardiac Cell', 'Cholesterol Biosynthesis', 'Complement and Coagulation Cascades' or 'Diabetes pathways'.

Gene Ontology
The Gene Ontology (GO) annotations were downloaded from the AmiGO 2 browser. The Gene Ontology (GO) project provides information about gene product function using ontologies to represent biological knowledge in the form of three classes: (1) molecular functions, (2) the biological processes these contribute to and (3) the cellular locations where these occur (cellular components). We first collected all GO terms associated with each of the candidate genes (Supplementary Table 12) and performed (1) enrichment analysis to highlight the over-represented categories, as described above for the Mouse Phenotype analysis and (2) to score the candidate genes based on their mapping to CAD-relevant GO categories, whereas CAD-relevance was determined similarly, as for the Biochemical Pathways described above and
summarized in Supplementary Table 13. Each candidate gene received a score of 1, if it was mapped to at least one CAD-relevant GO term, such as ‘GO:0002687-positive regulation of leukocyte migration’, ‘GO:0004465-lipoprotein lipase activity’, ‘GO:0005509-calcium ion binding’ or ‘GO:0006955-immune response’.

**Tissue-specific regulatory networks in human and key driver analysis**

In order to determine the gene regulations of CAD genes, we retrieved human tissue-specific Bayesian network models constructed from transcriptomic and genetic datasets from multiple human and mouse studies, including adipose tissue, liver, blood, brain, kidney, islet, and muscle. The network topology of these tissue-specific Bayesian networks defines a partitioned joint probability distribution over all genes, capturing detailed gene-gene regulatory relationships of CAD genes.

We applied a previously developed key driver analysis algorithm of gene-gene interaction networks to CAD datasets in order to identify the key regulatory genes. We used Bayesian gene networks due to their ability to capture detailed gene-gene relationships with causal implications. In particular, we analyzed networks from several tissues, including adipose tissue, liver, blood, brain, kidney, islet, and muscle. We explored the gene regulatory patterns of the detected CAD suspected genes above (Protein-coding + eQTL) in human tissue-specific Bayesian networks. The key driver genes for the CAD subnetworks are listed in Supplementary Table 6.

**Identifying mouse phenotypes from Systems Genetics Database (SGD)**

The Hybrid Mouse Diversity Panel Systems Genetics Database (http://systems.genetics.ucla.edu) contains multi-scale (clinical trait, transcriptomic, proteomic, metabolome, and epigenomic) data for several mouse crosses and a panel of 100 inbred strains of mice known as the Hybrid Mouse Diversity Panel. Among the clinical traits are atherosclerosis, heart failure, plasma lipids, and insulin resistance.
Materials and Methods References