Original Research

Systems Genetics Analysis of Genome-Wide Association Study Reveals Novel Associations Between Key Biological Processes and Coronary Artery Disease


Objective—Genome-wide association studies have identified multiple genetic variants affecting the risk of coronary artery disease (CAD). However, individually these explain only a small fraction of the heritability of CAD and for most, the causal biological mechanisms remain unclear. We sought to obtain further insights into potential causal processes of CAD by integrating large-scale GWAS with expertly curated databases of core human pathways and functional networks.

Approaches and Results—Using pathways (gene sets) from Reactome, we carried out a 2-stage gene set enrichment analysis strategy. From a meta-analyzed discovery cohort of 7 CAD genome-wide association study data sets (9889 cases/11089 controls), nominally significant gene sets were tested for replication in a meta-analysis of 9 additional studies (15502 cases/55730 controls) from the Coronary Artery Disease Genome wide Replication and Meta-analysis (CARDIoGRAM) Consortium. A total of 32 of 639 Reactome pathways tested showed convincing association with CAD (replication \( P < 0.05 \)). These pathways resided in 9 of 21 core biological processes represented in Reactome, and included pathways relevant to extracellular matrix (ECM) integrity, innate immunity, axon guidance, and signaling by PDRF, NOTCH, and the transforming growth factor-β/SMAD receptor complex. Many of these pathways had strengths of association comparable to those observed in lipid transport pathways. Network analysis of unique genes within the replicated pathways further revealed several interconnected functional and topologically interacting modules representing novel associations (e.g., semaphoring-regulated axonal guidance pathway) besides confirming known processes (lipid metabolism). The connectivity in the observed networks was statistically significant compared with random networks \(( P < 0.001 \)). Network centrality analysis (degree and betweenness) further identified genes (e.g., NCAM1, FYN, FURIN, etc.) likely to play critical roles in the maintenance and functioning of several of the replicated pathways.

Conclusions—These findings provide novel insights into how genetic variation, interpreted in the context of biological processes and functional interactions among genes, may help define the genetic architecture of CAD. (Arterioscler Thromb Vase Biol. 2015;35:00-00. DOI: 10.1161/ATVBAHA.115.305513.)

Key Words: coronary artery disease | pathway analysis

Meta-analysis of genome-wide association studies (GWAS) involving tens of thousands of subjects have provided a wealth of new information on the genetic basis of coronary artery disease (CAD), yet common susceptibility variants with achieved genome-wide significance explain only a small fraction of the heritability of CAD \(( \approx 10.6\% ) \). It has been proposed that much of the residual genetic risk may be attributable to rare variants with large effect. However, recent simulation, exome sequencing, and fine mapping studies of established GWAS loci support the hypothesis that joint contributions from common variants with modest effects are likely to account for a sizeable fraction of the missing heritability of complex diseases. It is likely that many more common variants are linked to CAD but have not achieved genome-wide significance in GWAS because of small effect size or lower allele frequency.
and insufficient sample size. However, based on the premise that clinically informative polymorphisms related to complex disease occur in systems of closely interacting genes,8 even weakly associated variants may provide important information about the biological basis of disease when such variants cluster within a common functional module or pathway. One common approach for pathway-based analysis of genomic data is gene set enrichment analysis (GSEA), originally developed and extensively used for the analysis of gene expression data.9 In 2007, Wang et al10 described a modified version of the GSEA, designed to analyze genome-wide single-nucleotide polymorphism (SNP) associations rather than gene expression data. Since then, several other GSEA methods have been developed for this purpose.11–15 The common goal of these analytic algorithms is to identify a subset of genes whose variants collectively demonstrate strong association with a trait of interest even if the component SNPs individually exhibit relatively modest or nonsignificant association. Importantly, pathway analysis can also place the set of validated SNPs for a trait of interest into a broader and clearer biological context. A natural extension of this list-based pathway approach is the interrogation of molecular networks to unravel the architecture underlying complex diseases. A molecular network is based on interactions among biomolecules (genes, proteins, metabolites, etc.), where such interactions can take various forms (protein–protein interactions, coexpression, gene regulation, functional interactions, etc.). Efforts at the characterization of disease-associated genes reveal that genes associated with the same or similar disorders tend to occupy similar neighborhoods in molecular networks through physical or functional modules.16,17 Furthermore, the study of network topology suggests that key disease-related genes differ from other genes in terms of their network connectivity and network centrality properties.17 Finally, molecular networks provide 2 distinct enhancements over traditional pathway-based approach—(1) they provide additional information on interactions among gene subsets within a given pathway, and (2) they allow for the identification of interactions between components of different biological pathways. Through these analyses, 1 is able to draw a clearer picture of the functional connectivities that influence pathway functions, and how multiple pathways may interact with 1 another to influence a phenotype.

Several studies have applied molecular networks for generating insights from GWAS data3,18–20 in disorders, such as schizophrenia, multiple sclerosis, and prostate cancer. However, most of these approaches have relied mainly on protein–protein interaction networks, thereby missing the rich mechanistic information available from traditional biological pathway repositories and networks based on functional interactions. In this study, we have coupled the advantages of a well-curated biological pathway repository with a similarly curated functional interaction network to identify mechanism-based processes that may underlie the genetic architecture of CAD. First, to identify novel associations between established biological mechanisms and CAD, we have carried out a 2-stage pathway-based GSEA analysis of 16 GWAS data sets for CAD using the i-GSEA4GWAS (http://gsea4gwas.pysch.ac.cn/inputPage.jsp) tool15 and the Reactome pathway database.21 Collectively, these GWAS include >25,000 subjects with CAD and >66,000 controls. We have then taken the replicated pathways as a starting point to explore functional interactions within and between pathways via interrogation of molecular interacting networks. Finally, we have characterized the CAD-associated genes based on their topological properties within these networks as a way of prioritizing gene candidates for functional follow-up studies.

**Materials and Methods**

Materials and Methods are available in the online-only Data Supplement. Briefly, using pathways (gene sets) from Reactome, we first carried out a 2-stage gene set enrichment analysis strategy. From a meta-analyzed discovery cohort of 7 CAD GWAS data sets (9889 cases/11089 controls), nominally significant gene sets were tested for replication in a meta-analysis of 9 additional studies (15,502 cases/35,730 controls) from the CARDIoGRAM Consortium (Table I). Genes from the replicated pathways were then mapped onto well-curated interaction networks.

**Results**

**Significant Pathways**

A total of 85 of the 639 Reactome pathways tested in stage 1 achieved a gene set enrichment \( P < 0.05 \) at a false discovery rate (FDR) < 0.25. Thirty-two of these 85 pathways were further replicated in stage 2 at a nominal \( P < 0.05 \) (Table 2). When the replicated pathways were compared with the full pathway content of Reactome, at least 1 replicated from 9 of the 21 core Reactome-defined biological processes. These included the core processes of metabolism, signal transduction, developmental biology, ECM organization, immune system, metabolism of proteins, cell–cell communication, transmembrane transport of small molecules, and gene expression (Figure 2). Because of the hierarchical organization of Reactome pathways, several replicated pathways were nested within larger gene sets, either completely or partially (Figure II in the online-only Data Supplement). This hierarchical structure enabled us to identify instances of pathway selectivity—for example, although the CRMPS in SEMA3A signaling, Sema4D in semaphorin signaling, and Sema3A PAK-dependent axon repulsion pathways all nested completely within the Semaphorin Interactions pathway, only the former was significantly replicated (\( P < 0.001 \)), whereas the latter 2 pathways were not. To put the identified pathways in a broader context, we have also listed the nonreplicated pathways that share similar levels of hierarchy as the replicated pathways in Table I (online-only Data Supplement).
About a third of the 32 replicated pathways were also significant in stage 2 \((P<0.05)\) after correcting for linkage disequilibrium between the SNPs, by analyzing SNPs pruned genome-wide at either \(r^2>0.5\) or \(r^2>0.2\) (Table II in the online-only Data Supplement). The pathways that were in common to all 3 pruned and unpruned SNP analyses were Toll receptor cascades, degradation of the ECM, lipid digestion, mobilization, and transport, and lipoprotein metabolism. Although the association of these pathways may be of higher confidence, pruning of SNPs may also lead to loss of power because of significant reduction in SNP number (5% to 15% of unpruned SNPs) and to the fact that the pruning was agnostic to the actual CAD SNP association \(P\) values. Hence, for downstream gene and network analyses we chose to use the full set of 19 pathways that replicated with the unpruned list of SNPs.

Finally, we examined the possible effect of LD among genes leading to inflated significance scores for the replicated pathways by considering the extent of LD among the gene-tagging (best scoring) SNPs for all genes in a pathway. The extent of LD among the most significant SNPs was found to be minimal. Specifically, of all the SNPs tested, we found only 2 SNP pairs with an \(r^2>0.8\), observed across 3 pathways. Even at the more permissive \(r^2\) threshold of 0.2, only 4 SNP pairs were observed across 5 pathways (Table III in the online-only Data Supplement).

**Gene and Pathway Prioritization**

The 32 replicated pathways contained a total of 770 unique genes that were taggable by at least 1 SNP (no SNP tags were available for 83 genes). Figure SIII (online-only Data Supplement) summarizes the proportion of genes within the replicated pathways that were associated with CAD. All replicated pathways contained \(\geq 50\%\) genes above the significance threshold (range, 50.0%–92.3%), confirming that the pathway findings were driven by the combined contributions of multiple genes in each pathway and not because of large effects from a small minority of genes. For comparison purposes, we also analyzed a synthetic pathway derived from genes within the CARDIoGRAM loci reaching genome-wide significance. This synthetic pathway contained the second highest proportion of genes reaching the significance threshold.

**Network Analysis**

**Statistical Evaluation of Network**

A total of 770 genes from the replicated pathways were mapped to the InWeb PPI network and the observed network connectivity parameters (degree, and number of edges) compared with random networks of similar size and degree distribution. A network of direct interactions could be created with 620 genes (assuming a minimum interaction size of 2 genes). The resulting network (Figure SIV) was significantly different with respect to random networks; thus there were 3726 direct edges in the network compared with only 1548 edges expected by chance \((P<0.001)\), and the observed average connectivity per gene (degree of gene) was 12, compared with an expected 5.8 from random networks \((P<0.001)\). These results indicate that the networks constructed from the replicated pathway genes are probably not because of chance.

**Table 1. Demographics of Discovery and Replication Cohorts**

<table>
<thead>
<tr>
<th>GWAS Data Set</th>
<th>No. of Cases/Controls</th>
<th>Age (mean±SD)</th>
<th>% Men Cases/Controls</th>
<th>% MI Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stage 1 studies</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OHGS_A</td>
<td>921/994</td>
<td>48.2±7.0/74.9±4.9</td>
<td>78.1/54.6</td>
<td>54.6</td>
</tr>
<tr>
<td>OHGS_CCGB-B</td>
<td>2688/1819</td>
<td>49.8±7.7/77.4±5.4</td>
<td>75.1/49.0</td>
<td>59.8</td>
</tr>
<tr>
<td>DUKE_2</td>
<td>1200/648</td>
<td>56.7±9.7/63.3±8.7</td>
<td>69.4/42.0</td>
<td>48.0</td>
</tr>
<tr>
<td>GerMIFs I</td>
<td>875/1644</td>
<td>50.3±7.8/62.6±10.0</td>
<td>50.6/49.2</td>
<td>100</td>
</tr>
<tr>
<td>GerMIFs II</td>
<td>1222/1298</td>
<td>51.4±7.5/51.2±11.9</td>
<td>66.9/51.7</td>
<td>100</td>
</tr>
<tr>
<td>GerMIFs III (KORA)</td>
<td>1157/1748</td>
<td>58.6±8.7/55.9±10.7</td>
<td>79.9/51.1</td>
<td>100</td>
</tr>
<tr>
<td>WTCCC</td>
<td>1926/2938</td>
<td>49.8±7.7/N/A</td>
<td>79.3/50.0</td>
<td>71.5</td>
</tr>
<tr>
<td><strong>Total stage 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9889/11089</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Stage 2 studies</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADVANCE</td>
<td>278/312</td>
<td>45.8±6.2/45.3±5.7</td>
<td>42.1/41.0</td>
<td>50.4</td>
</tr>
<tr>
<td>CADomics</td>
<td>2078/2952</td>
<td>60.8±10.1/55.3±10.8</td>
<td>78.1/49.5</td>
<td>58.3</td>
</tr>
<tr>
<td>CHARGE</td>
<td>2287/22024</td>
<td>60.0±9.9/63.1±8.0</td>
<td>66.6/40.4</td>
<td>48.0</td>
</tr>
<tr>
<td>deCODE CAD</td>
<td>6640/27611</td>
<td>74.8±11.8/53.7±21.5</td>
<td>63.7/38.1</td>
<td>54.7</td>
</tr>
<tr>
<td>LURIC/AtheroRemo 1</td>
<td>652/213</td>
<td>61.0±11.8/58.3±12.1</td>
<td>79.7/54.0</td>
<td>71.9</td>
</tr>
<tr>
<td>LURIC/AtheroRemo 2</td>
<td>486/296</td>
<td>63.7±9.4/56.4±12.7</td>
<td>76.6/51.4</td>
<td>79.0</td>
</tr>
<tr>
<td>MedStar</td>
<td>874/447</td>
<td>48.9±6.4/59.7±8.9</td>
<td>67.0/45.4</td>
<td>48.1</td>
</tr>
<tr>
<td>MiGen</td>
<td>1274/1407</td>
<td>42.4±6.6/43.0±7.8</td>
<td>62.8/60.1</td>
<td>100</td>
</tr>
<tr>
<td>PennCATH</td>
<td>933/468</td>
<td>52.7±7.6/61.7±9.6</td>
<td>76.3/48.1</td>
<td>50.3</td>
</tr>
<tr>
<td><strong>Total stage 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>12501/55730</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Total stages 1 and 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>25491/66819</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CAD indicates coronary artery disease; and GWAS, genome-wide association study.
Although this PPI-based analysis provided confidence that
the networks derived from the replicated pathway genes are
unlikely to arise from chance, it allows only limited insights
into the various biological mechanisms impacted by these
pathways. Thus, to identify networks that contain more rel-

evante information on biological processes (including PPI),
the genes from the replicated pathways were mapped to a
functionally interacting network curated and maintained at
Reactome. A total of 733 genes could be mapped to the larger
network. This subnetwork was further clustered to reveal
within-network modules. Clustering resulted in the identifica-
tion of 17 clusters with 10 clusters containing >10 gene mem-
bers (Figure 3; Table SIV). Within each cluster, a diverse array
of interactions (reactions, complex formation, activation, etc.)
was represented by the edges connecting the genes (nodes),
as exemplified in Figure V (online-only Data Supplement)
for the genes in clusters 8 and 9. We also observed consid-

erable interconnectivity between the clusters; for example, the
links between cluster 4 and other clusters are highlighted in
Figure 3 (additional intercluster connectivities for each of the
remaining clusters are shown in Figure VI in the online-only
Data Supplement). Enrichment analysis within each cluster
using Gene Ontology (http://www.geneontology.org/) identi-
fied several cluster-specific overrepresentations of biological
processes, as further highlighted in Figure 3. The following
are some notable examples of functional enrichment within
the clusters (FDR<0.001): innate immunity (clusters 1 and
Gene and Pathway Prioritization Based on Network Topology

Network topology provides vital information toward the understanding of network architecture and performance and allows for the prioritization of genes based on their topological characteristics within the network. Thus, we interrogated the topological properties of the networks derived from the replicated pathways. Specifically, we investigated 2 key node centrality measures, namely degree and betweenness because of their reported significance in biological networks as drivers for gene/protein essentiality (see online-only Data Supplement for additional information on degree and betweenness). For this purpose, the replicated pathways were first converted into Reactome functional interaction networks (conversion was successful for 29 pathways, with the exclusion of collagen formation, metabolism of polyamines, and organic cation–anion zwitterion transport pathways) and subsequently analyzed for the above 2 node centrality measures. Figure 4 depicts the betweenness centrality measures for a merged network derived from 2 pathways related to cell–cell interactions (neural cell adhesion molecule [NCAM] signaling for neurite outgrowth and CRMPs in Sema3a signaling). In this network, the NCAM1 and Fyn proteins display large betweenness centrality and act as bridges connecting multiple other proteins in the network. Some additional genes with GWAS association $P < 0.001$ that occupy potentially critical positions (betweenness $>100$) in a subset of the replicated pathways include FURIN (component of degradation of ECM, ECM organization, signaling by NOTCH1 pathways), MMP1 (degradation of ECM and ECM organization pathways), and RPS6KA5 (Toll receptor cascades and growth factor signaling pathways).
NCAM signaling for neurite outgrowth pathways). Results for the remaining pathways are shown in Figure VII and Table VI (online-only Data Supplement).

**Discussion**

Despite the recent successes of large GWAS meta-analyses, the genetic architecture of CAD remains poorly understood and the identified loci explain a small proportion of genetic risk. By integrating GWAS data with expertly curated databases of core human pathways as well as gene and reaction-based functional networks, we sought to obtain novel insights into the potential causal processes of coronary atherosclerosis. In addition, the large size of the discovery population and replication sample (25,000 CAD cases and 66,000 controls) and the 2-step discovery-replication strategy increases confidence in the results. This analysis implicates 32 core human pathways representing 9 distinct biological processes as being most etiologically relevant to CAD.

Notably, many replicated pathways from the 2-stage GWAS analysis strategy converged on processes regulating cellular growth, migration, and proliferation, such as the signaling by transforming growth factor-β receptor and signaling by PDGF, pathways previously intensively investigated for their functional role in coronary atherosclerosis. By combining GWAS-based findings with such a priori information, we obtained evidence that genetic variation in a critical number of genes representing these pathways contribute to the heritability of CAD. Moreover, these data support hypotheses that alterations in these pathways are potentially causally related to CAD. Specifically, transforming growth factor-β is known to control cell proliferation, cell migration, matrix synthesis, wound contraction, calcification and the immune response, all of which are major components of the atherosclerotic process. PDGF is expressed in every cell type of the atherosclerotic arterial wall, as well as in infiltrating inflammatory cells and plays a key role in the migration of vascular smooth muscle cells from the media into the intima and their subsequent proliferation. Although both pathways have been studied in animal models, animal data are often conflicting or inadequate and there are no data related to modulation of...
these pathways in humans. Several pathways related to the integrity of the ECM were also highly significant, including ECM organization, degradation of the ECM, and cell ECM interactions. The ECM is responsible for maintaining not only the structural integrity of vessel wall plaques but also participates in several key events, such as cell migration, lipoprotein retention, and thrombosis that are critically linked to plaque stability.25

Two of the axon guidance pathway subclasses, such as CRMPs in Sema3 signaling, and NCAM signaling for neurite outgrowth also replicated. The axon guidance pathways modulate diverse biological phenomena, including cellular adhesion, migration, proliferation, differentiation, survival, and synaptic plasticity through the participation of highly conserved families of guidance molecules, including netrins, slits, semaphorins, and ephrins, and their cognate receptors.26 Neural guidance cues such as netrin-1 and semaphorins have important roles outside the nervous system. Oksala et al27 provide compelling evidence that netrin-1 is secreted by macrophage foam cells in atherosclerotic plaques and acts to inhibit emigration of these cells out of lesions by causing dysregulation of the actin cytoskeleton. Wanschel et al28 reported that NTNI is downregulated in atherosclerotic plaques and its expression correlates negatively with inflammatory markers and M2 signals. Like netrin-1, semaphorin 3A, encoded by SEMA3A, one of the top-ranked genes in this analysis, is also expressed in coronary artery endothelial cells and potently inhibits chemokine-directed migration of human monocytes.29,30 This study also provides further supportive evidence for a causal role of innate immunity in atherosclerosis or plaque rupture with significant pathways, including both toll receptor cascades and initial triggering of complement. Innate immune responses mounted by macrophages and other immune cells recruited to the arterial wall in response to an inflammatory challenge have a major role in the initiation of atherosclerosis.31

An important advance encompassed in the current work is our further examination of the topological characteristics of genes comprising the replicated gene sets and the potential implication of topology on biological function. Specifically,
we applied the Reactome FI tool to identify gene sets related to biological processes, such as innate immunity, cell adhesion, and lipid metabolism that were further reorganized into functionally interacting networks and subnetwork clusters demonstrating a high degree of interconnectedness. Network clustering, followed by pathway enrichment analysis on the identified clusters via Gene Ontology, generated new insights on interrelationships among the enriched pathways, not available through our initial traditional gene set analysis. For example, whereas the lipid-metabolizing genes were largely concentrated in a single cluster (cluster 8), genes related to innate immunity were, by contrast, distributed within 3 separate clusters (clusters 0, 1, and 4), along with other biological processes, highlighting the possibility of extensive interactions among these processes. Finally, through analysis of such networks, we were further able to evaluate the possible criticality of genes in network function, based on the degree and betweenness centrality properties of the network genes.

Collectively, these additional analytic approaches provide important insights into the interrelationships among genes that are not usually available through conventional gene set enrichment analysis, and could assist in the formation of testable hypotheses on areas of robustness and vulnerability in functional networks otherwise not intuitively evident. For example, topological analysis implicated a potential role for the axonal growth related pathways in CAD with NCAM1 being a major hub in a network, including plexins (PLXNA1 and PLXNA2), neuropilin-1, as well as adhesion molecules (CNOT2) and several members of the collagen family relevant to the ECM of the vessel wall (Figure 4). These data support the concept that neuronal guidance cues have important roles in both arteriogenesis and atherosclerosis by regulating macrophage retention in plaques.27,29,30 Other studies demonstrate that semaphorin 3A and its receptors, neuropilin-1 and -2, plexins A1/A2/A3 are highly expressed in human monocye-derived macrophages and play a role in induction of macrophage apoptosis.34

Despite these plausible observations, we are cognizant that betweenness is but only one of several network centrality measures that could play critical roles in network function. Because both fields of network biology and network pharmacology are currently evolving, our findings should be considered more as hypotheses-generating rather than conclusive evidence of the importance of a gene or 1 pathway over another. Functional testing is necessary as the next step, and can take several forms, including (1) overexpression or knockdown of medium to high betweenness genes in target pathways (eg, NCAM1, FYN, for the network in Figure 4) in CAD-relevant cell models (eg, human coronary artery endothelial or smooth muscle cells, macrophages, etc.) and to interrogate their effects on cell function (cell migration, lipid accumulation, etc.); (2) testing the effects of candidate genes (eg, NCAM1 and FURIN) in knockout or overexpression mouse models (generated by somatic manipulation or transgene creation) on lesion formation (similar to studies on candidate GWAS genes for lipoprotein metabolism); (3) statistical epistasis analysis, limited to genes within a replicated pathway, to uncover functionally important interactions underlying the genetic basis of atherosclerosis, and (4) prioritizing gene products from replicated pathways based on the availability of pharmacological agents against them, and testing these for potential benefits in animal models of atherosclerosis (successfully demonstrated in identification of memory-modulating drugs).39 We hope our approach stimulates extensive further discussion on how to experimentally interrogate CAD related networks and pathways.

We acknowledge potential caveats pertaining to this study. First, the number of pathways identified and replicated was modest but the pathways are biologically plausible. In the discovery analysis, 85 of the 639 (13%) pathways tested were significant at \( P<0.05 \) (and FDR <25%) with at least 50% of the genes in any given pathway being individually significant at a \( P<0.05 \). A total of 32 of these 85 (37%) pathways, achieved replication, a number somewhat lower than expected (75%) given the FDR threshold used in the discovery phase to select pathways for testing in the replication sample. This may reflect the less stringent criteria for age of onset of CAD cases applied in some of the replication studies as well as study-specific differences in inclusion/exclusion criteria and adjudication of outcomes, leading to increased sample heterogeneity.1 Our study also highlights several generic issues that currently impose limitations on the conduct and interpretation of pathway analyses.40 Some of these issues pertain to (1) the mapping of SNPs to genes, (2) choosing the optimum pathway analysis tool for GWAS, (3) consequences of the permutation scheme used in i-GSEA4GWAS, and (4) the effects of inter-SNP linkage disequilibrium on pathway analysis results. An additional caveat is the potential for bias in the network and topological analyses because of limitations in the extent and type of experimental data available in the source databases. We have provided a further detailed discussion of issues related to pathway and network analysis in the Results section of the online-only Data Supplement.

This is an area of emerging methodology and different approaches can yield complementary findings. Our findings extend gene-centric verification of CAD GWAS loci and those recently reported by CARDIoGRAM+C4D, applying Ingenuity network analysis only on the top 239 candidate genes.2 In another recently published study, based on this large-scale meta-analysis of GWAS studies for CAD, we used a different approach.42 Rather than a location-based approach to map SNPs to genes, we used eQTL data from CAD-related tissues and primary cells to link CAD SNPs to their empirically defined target genes. We then created data-driven, tissue-specific gene expression networks from a multitude of human and mouse experiments.42 These networks relied heavily on available gene expression data and did not involve other types of interactions, such as protein–protein interactions or biochemical reactions. In contrast, this analysis is based on gene-to-SNP mapping methods for gene set enrichment rather than eQTL data and our analysis of the topological relationships among genes in the filtered, replicated pathways using Reactome FI and pathway interaction database cover a more extensive array of molecular interactions, thus revealing important aspects that we failed to capture from the gene expression-based networks. It is encouraging that these 2 approaches have yielded consistent results in terms of core processes related to lipid
metabolism, immune system, Notch-HLH transcription and PPAR signaling. However, here we have identified additional biologically relevant pathways, including ECM integrity, transforming growth factor-β signaling and axon guidance, the latter being of particular interest given recent laboratory findings.27–30,32–34 Many of these pathways had strengths of association comparable to those observed in known pathways related to lipoprotein metabolism.

The findings of this extensive but preliminary analysis do not imply causality. However, the use of the integrative approach in elucidating the genetic bases of disease has been demonstrated by studies in several complex phenotypes. For example, in an investigation of the WTCCC Crohn disease GWAS data set, only 3 genes at 2 loci showed GWAS significant signals but pathway analysis carried out by Wang et al11 identified the 20 gene IL-12/IL-23 pathway to be associated with Crohn disease that remained significant even when the 2 original loci were removed.41 In a similar vein, Holmans et al44 provided supporting evidence for the immunogenetic origins of Parkinson disease by identifying the regulation of leukocyte/lymphocyte activation and cytokine-mediated signaling as conferring increased susceptibility to Parkinson disease, although none of the SNPs linked to genes within these pathways had achieved GWAS significance. On the contrary, pathway analysis studies have had little success in generating new biological insights for other disorders, including type 2 diabetes mellitus. Because of this variability, extensive mechanistic biological insights for other disorders, including type 2 diabetes mellitus, may facilitate the development of novel testable hypotheses that could ultimately improve our understanding of atherosclerosis.

Appendix
From the Program in Cardiovascular and Metabolic Disorders (S.G.) and Centre for Computational Biology (S.G.), Duke-NUS Graduate Medical School, Singapore, Singapore; Department of Cardiovascular and Metabolic Research, Biomedical Biotechnology Research Institute, North Carolina Central University, Durham (S.G., J.V.); Department of Cardiovascular Sciences, University of Leicester, Glenfield Hospital, Leicester, UK (C.P.N., N.J.S.); Institut für Integrative und Experimentelle Genomik (IIEG), Universität zu Lübeck, Lübeck, Germany (C.W., J.E.); DZHK (German Research Centre for Cardiovascular Research), Partner Site Hamburg, Kiel, Lübeck, Germany (C.W., J.E.); Broad Institute of Harvard and MIT, Cambridge, MA (A.V.S., S.K.); Department of Integrative Biology and Physiology, University of California, Los Angeles (V.-P.M., X.Y.); Atherogenomics Laboratory (M.N., R.M.P.), John and Jennifer Ruddy Canadian Cardiovascular Research Centre (A.F.R.S., R.M.P.), and Division of Cardiology (R.M.P.), University of Ottawa Heart Institute, Ottawa, Canada; Clinic for General and Interventional Cardiology, University Heart Center Hamburg, Germany (S.B.); National Heart, Lung, and Blood Institute’s Framingham Heart Study, MA (C.O.D.); Mannheim Institute of Public Health, Social and Preventive Medicine, University of Heidelberg, Germany (W.M.); Synlab Academy, Mannheim, Germany (W.M.); Science Center, Tampere University Hospital, Tampere, Finland (R.L.); Cardiovascular Research Institute, Washington Hospital Center (S.E.E.); Department of Medicine, Duke University Medical Center, Durham, NC (S.H.S., C.B.G.); Cleveland Clinic, OH (S.L.H.); Cardiology Division, Center for Human Genetic Research (S.K.) and Cardiovascular Research Center (S.K.), Massachusetts General Hospital, Harvard Medical School, Boston; Cardiovascular Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia (M.P.R.); Department of Medicine, Stanford University School of Medicine, CA (T.Q., T.L.A.); National Institute for Health Research (NIHR) Leicester Cardiovascular Biomedical Research Unit, Glenfield Hospital, Leicester, United Kingdom (N.J.S.); Deutsches Herzzentrum München, Technische Universität München, Munich, Germany (H.S.); and DZHK (German Research Centre for Cardiovascular Research), Partner Site Munich Heart Alliance, Munich, Germany (C.P.N., H.S.).

Acknowledgments
We thank all the individuals who contributed to these multicentered studies. The full list of the investigators who are part of the CARDIoGRAM Consortium is listed in the S1 Material online-only Data Supplement. A full list of the investigators who contributed to the generation of the Welcome Trust data is available from http://www.wtccc.org.uk.

Sources of Funding
This work was supported by grants from the American Heart Association (AHA10SDG4230068 to S. Ghosh; 13POST17240095 to V.-P. Mäkinen; 13SDG17290032 to X. Yang); National Institutes of Health (1R21DK088319, 2P20MD00175-11 to S. Ghosh); (P01HL098055, P01HL076491, P20HL113452 to S.L. Hazen); (R01-HL095987 to S.H. Shah); (K23DK088942 to T.L. Assimes); Wellcome Trust (No. 076113 and No. 085475) British Heart Foundation (to N.J. Samani); European Union (No. 20168 to R. Laaksonen). Deutsche Forschungsgemeinschaft and the German Federal Ministry of Education and Research (BMBF) in the context of the German National Genome Research Network (NGFN-2 and NGFN-plus), the FP6 and FP7 EU funded integrated projects Cardiogenics (No. LSHM-CT-2006-037593) and ENGAGE (No. 201413), and the binational BMBF/ANR funded project CARDomics (No. 01KU0908A); Canadian Institutes of Health Research (No. MOP-2380941 to R. McPherson); (No. MOP28210, MOP77682 to A.F.R. Stewart and R. McPherson); Canada Foundation for Innovation (No. 11966 to RR, A.F.R. Stewart, and R. McPherson; Heart and Stroke Foundation of Canada (No. NA6001, No. NA6630 to R. McPherson).

Disclosures
None.
References


Genome-wide association studies have identified >45 loci associated with coronary artery disease (CAD) risk but provide limited insight into causal mechanisms. Furthermore, the identified signals explain little >10% of the predicted heritability of CAD. Part of this missing heritability is likely because many more common variants are linked to CAD but have not achieved genome-wide significance in genome-wide association studies because of small effect size or lower allele frequency and insufficient sample size. However, even weakly associated variants may provide important information about the biological basis of disease when such variants cluster within a common functional module or pathway. By integrating genome-wide association study data with extensive databases on core biological processes, we have identified novel biological pathways relevant to the pathogenesis of CAD. These findings provide new insight into how genetic variation, interpreted in the context of biological processes and functional interactions among genes, may help define the genetic architecture of CAD.
Systems Genetics Analysis of Genome-Wide Association Study Reveals Novel Associations Between Key Biological Processes and Coronary Artery Disease


Arterioscler Thromb Vasc Biol. published online May 14, 2015;
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/early/2015/05/14/ATVBAHA.115.305513

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/
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/
SUPPLEMENTAL MATERIAL

Acknowledgments

Supplementary Methods

Gene-set enrichment analysis methods

Calculation of enrichment score in i-GSEA4GWAS: For enrichment score calculations, i-GSEA4GWAS first represents a gene by the max(−log(p-value)) of all the SNPs mapped to that gene within the 100 kb window. Next, genes are ranked based on this association p-value. The ranked gene list is used to calculate the enrichment score (ES) of the pathways. The 'enrichment score' reflects the degree to which a pathway is over-represented at the extreme of the entire ranked list. The ES is then corrected by a ‘k/K factor’, where k and K are the proportions of significant genes in the pathway and among the total genes in GWAS, respectively (genes mapped with at least one of the top 5% of all SNPs are considered significant). A gene significance threshold is also calculated based on the number of genes that are mapped to one of the top 5% most associated SNPs out of all GWAS genes (in the current study, the threshold was 1.37). These corrected enrichment scores, also known as ‘significance-proportion based enrichment scores’, were used to calculate pathway association p-values and false discovery rates via SNP label permutation to correct for gene and gene set variation¹.

Pruning of SNPs in linkage disequilibrium (LD): To test the effect of adjusting for differences in linkage disequilibrium (LD) patterns across the genome on our GSEA results, we generated two lists of pruned SNPs (SNPs that are in approximate linkage equilibrium to each other) using two different LD thresholds. Pruning was done using genotype data (2,239,392 SNPs) of 60 unrelated CEU individuals from Hapmap2 release 22, build 36 (hg18), and was executed using the --indep-pairwise function in PLINK². SNPs with a pairwise $r^2>0.2$ or $r^2>0.5$ were removed within a sliding window of 200 SNPs, using window jumps of 5 SNPs. Pruning at $r^2>0.2$ or $r^2>0.5$ thresholds yielded 132,195 and 354,522 SNPs, respectively (5.4% or 14.6% of the full list of ~2.4 million unpruned SNPs).
**Determination of LD among genes in a pathway:** To test whether physical clustering of subsets of genes within each tested pathway could lead to biased estimates of pathway enrichment (due to linkage disequilibrium among the best scoring SNPs for those genes), we determined the degree of LD among the most significant SNPs mapped to genes within a pathway using the online SNP Annotation and Proxy Search tool (SNAP). We used the HapMap release 22 panel for CEU to calculate $r^2$ between the SNPs in gene sets that replicated.

**ReactomeFI tool for network analysis:** The gene sets employed in GSEA essentially represent ‘lists’ of genes that by themselves do not convey an understanding of the functional and topological relationships that may exist among such genes. However, such relationships are often crucial for biological function and can yield important information about the essentiality or vulnerability of different components of a pathway. The ‘Reactome-FIs’ (functional interactions) data set unites interactions from Reactome and other pathway databases, including KEGG, BioCyc, Panther, The Cancer Cell Map and Pathway Interaction Database (PID) with pair-wise interactions gleaned from physical protein–protein interactions in human and model organisms, gene co-expression data, protein domain–domain interactions, text mining and GO annotation. As of December 2012, the ‘Reactome-FIs’ network contains 209,988 functional interactions encompassing 10,956 proteins (excluding splice isoforms), reflecting 46% of SwissProt proteins and provides a pathway-informed data analysis system for high-throughput data analysis, leading to greater biological interpretability.

**Network clustering:** A larger network may be composed of smaller modules where the connections (edges) among members (nodes) within modules are dense but connections between modules are sparse. In technical terms, the modularity for a given division of a network can be defined as the fraction of edges within groups minus the expected fraction of such edges in an equivalent network with edges placed at random. Network clustering attempts to identify the modular substructure of networks by
using various algorithmic approaches (such as graph partitioning). The modules so identified may then be interrogated for specific functions which, if they exist, provides novel information about the functional architecture of the network and illuminate how modules may relate to one another. The clusters generated in our study were generated by spectral partitioning based methods detailed in Newman \textsuperscript{9}.

\textbf{‘Betweenness’ as a network property:} Betweenness of a node is defined as the fraction of the shortest paths between all pairs of nodes in a network that pass through that node and estimates the functional load through one node or link (assuming that information flows over a network predominantly via the shortest paths)\textsuperscript{11}. For three nodes, \(i, j, k\), the betweenness centrality measure \(C_{i}^{Btw}\) for node \(i\) is given by

\[
C_{i}^{Btw} = \sum_{j=1}^{N} \sum_{k=1}^{j-1} \frac{g_{jk}(i)}{g_{jk}}
\]

Where \(g_{jk}(i)\) is the number of shortest paths from \(j\) to \(k\) through \(i\) and \(g_{jk}\) is the total number of shortest paths between \(j\) and \(k\)\textsuperscript{12}.

\textbf{Supplementary Results:}

\textbf{Choice of window size for SNP-to-gene mapping:} The choice of the window was determined after comparing the results of SNP-to-gene mapping at 100 kb and 500 kb intervals. Specifically, the larger interval resulted in more instances of the same SNP being mapped to multiple genes, leading to a reduction in the number of unique SNP-gene mappings and contracting the range of best scoring SNP p-values. Some of these results are most likely spurious, essentially being consequences of a larger window allowance around the gene. Conversely, a larger proportion of unique SNP-gene mappings were
obtained with the smaller interval (Figure S1). Additionally, the 100 kb region has also been reported to encompass the majority of regulatory regions, such as cis-eQTLs\textsuperscript{13}.

**Frequency of gene occurrence among replicated pathways:** Pathway multiplicity score was plotted against the pathway’s association p-value (from iGSEA) to identify pathways that were enriched for genes with high gene multiplicity scores (Figure S3). Replicated but non overlapping pathways with the highest proportion of genes occurring in other replicated pathways are shown in the top right hand corner of the figure. In this analysis, pathways related to extracellular matrix organization and degradation, protein post-translational modification, amino acid metabolism, and lipid mobilization and transport were identified in the top quartile of pathway multiplicity scores.

**Prioritization of network nodes based on “betweenness” measure of centrality:** Referring to Figure 4, several of the larger red nodes with high “betweenness” measures are composed of complexes of various proteins e.g. ‘GBR2:pFAK bound to NCAM:Pfyn’ and ‘SEMA3A:NRP1:PlexinA:Fyn’, suggesting that these complexes play a key role in the function of this network. These red nodes contain at least one gene (in a complex) that was statistically significantly associated to CAD in the replication study (iGSEA score >1.37). Two nodes (’CREB1’ and ‘FGFR1c homodimer’) are shown in yellow indicating that the iGSEA association statistic was less than 1.37 for these genes. In addition, the network contains some nodes colored in white, indicating that no SNP could be mapped to the genes in these nodes. However, these genes are included in the network to maintain network continuity.

**Generic issues with pathway analysis:** First, consensus is lacking regarding the optimal strategy for pathway-based approaches despite their enormous potential to extend the single SNP association tests of conventional GWAS analyses and to lead to the formulation of new biological hypotheses\textsuperscript{14}. In the absence of such consensus, we chose to present the pathway results of a single GSEA algorithm, iGSEA4GWAS, using a single pathway database, REACTOME. This decision was made following
preliminary studies, comparing several tools including GSA-SNP and MAGENTA and additional pathway repositories including KEGG and Biocarta. The results of this comparative analysis will be the topic of a separate manuscript. Second, despite its improved algorithmic approach to identify significantly associated pathways, iGSEA4GWAS is unable to compute exact p-value when these values are <0.001 because the program is limited to the generation of 1000 permutation datasets to derive the null distribution. Thus, pathways with p <0.001 are hard to prioritize based solely on their level of significance. Third, the 100 kb window commonly used to map a SNP to a gene may not adequately account for the potential effects of more distal regulatory SNPs. However, a recent investigation of the distribution of association signals in 7 GWAS datasets detected significant enrichment up to 40 kb upstream and downstream of protein coding genes, suggesting that a 100 kb window is most likely adequate to capture the majority of cis-regulatory signals\textsuperscript{15}. Fourth, the extent to which linkage disequilibrium between SNPs influences pathway analysis results remains an open question in the field. While it is easy to see how correlations among SNPs in LD can influence pathway analysis methods that utilize multiple SNPs to arrive at a gene significance score\textsuperscript{16-18}, the impact of LD on methods, such as i-GSEA4GWAS, that choose a single best SNP to tag a gene and correct for number of SNPs tested per gene is less clear. Pruning of SNPs (removal of SNPs above an LD-correlation threshold) has been proposed as a way to reduce the impact of LD on gene scores\textsuperscript{1}. However, pruning also leads to significant loss of information and reduces power for pathway analysis methods that depend on the accumulated evidence from all genes within a pathway to estimate pathway significance. Nevertheless, we also examined the effect SNP pruning on replication of the 85 pathways identified as nominally significant in the Discovery cohort. Generally, pruning reduced the number of testable SNPs to 5-15% of the original number, and only a small subset of the original 32 replicated pathways could be identified with the pruned datasets. We believe this result to be most likely a consequence of reduced information content of the pruned datasets, so that only the most significantly associated pathways could be
identified. Hence, we chose to identify candidate associated pathways (accepting any inherent biases that might be present) and then to use additional methods to reduce the number of false positives.

Lastly, LD among genes is another very relevant factor for pathway analysis, since inflation of pathway association score can occur if the index SNPs used to derive gene-based association p-values for proximally located genes are also in LD with each other. However, our LD analysis among the SNPs that were mapped to replicated gene sets suggests that over-inflation is unlikely to be a major concern in the current analysis given only 2 pairs of SNPs in 3 pathways were found to be in high LD (r^2 > 0.8).

Additionally, pathways with known instances of positionally clustered gene components (e.g. the HLA gene cluster or the chemokine gene cluster) were not significant in our analysis.

**Supplementary References:**

**References**


2. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, and Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81 (3):559-575


   Ref Type: Generic


Supplementary Figure SI a-d: Comparison of SNP-to-gene mapping using 100 or 500 kb windows around genes.

### a.

<table>
<thead>
<tr>
<th></th>
<th>100kb</th>
<th>500kb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of genes mapped</td>
<td>1834</td>
<td>1850</td>
</tr>
<tr>
<td>Number of different SNPs</td>
<td>1580</td>
<td>1177</td>
</tr>
<tr>
<td>Number of common SNPs</td>
<td></td>
<td>434</td>
</tr>
</tbody>
</table>

Summary of data used for the comparative analysis. Analysis was conducted on the set of genes present in pathways identified as nominally significant in the Discovery cohort (p<0.05).

### b.

Bivariate plot, showing for each gene (blue dot) the total number of mapped SNPs. Expectedly, the number of mapped SNPs per gene is greater for the 500 kb window than for 100 kb.

### c.

Histograms showing the distribution of SNPs mapping to 1, 2 or more genes when using a 500 vs. 100 kb window. The number of SNPs mapping to >1 gene is greater in the 500 kb window; conversely more unique SNP-to-gene mapping is observed with the 100 kb window.

### d.

Boxplot showing the range of p-values observed for the best-scoring SNPs when using a 500 kb vs. a 100 kb window. The range is the expected 0-1 for the 100kb but much contracted for 500kb.
Supplementary Figure SII: Degree of overlap among the 32 replicated pathways

Overlap comparison among the replicated pathways. Pathways are listed in rows and again in columns (in same order) and overlap among the pathways is computed as a fraction of 1.0. The extent of overlap is always expressed as the overlap of a pathway in column to the pathways in rows. The degree of overlap is represented in a yellow-red color scale, with higher overlaps indicated in shades of red.
Supplementary Figure SIII: Distribution of gene scores for replicated pathways
Supplementary Figure SIV: Protein-protein interaction network from replicated pathway genes

<table>
<thead>
<tr>
<th>PARAMETER</th>
<th>OBSERVED</th>
<th>EXPECTED</th>
<th>P_VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct Edge Count</td>
<td>3726</td>
<td>1548.137</td>
<td>0.001653</td>
</tr>
<tr>
<td>Seed Direct Degrees Mean</td>
<td>12</td>
<td>5.792533</td>
<td>0.001653</td>
</tr>
<tr>
<td>Seed Indirect Degrees Mean</td>
<td>587.8232</td>
<td>418.2798</td>
<td>0.001653</td>
</tr>
<tr>
<td>Seed CI Degrees Mean</td>
<td>6.464137</td>
<td>5.934328</td>
<td>0.001653</td>
</tr>
</tbody>
</table>
Supplementary Figure SV: Types of functional interactions observed within network clusters
Supplementary Figure SVI: Inter-cluster connectivities among networked genes from the replicated pathways.
Supplementary Figure SVII

**Topology based network analysis of replicated pathways.** “Betweenness” measure of node centrality was determined for genes in the replicated pathways. Pathways were first represented as Reactome functional interaction networks in Cytoscape. Betweenness centrality in the network nodes (genes) was calculated via the Centiscape 2.0 Cytoscape plugin. In the following diagrams, genes are color coded by their Stage 2 analysis p-values (deep red, p<0.001; lighter red, 0.001<p<0.01; lightest red, 0.01<p<0.05; white, p>0.05) and sized by their betweenness score. The gene names and betweenness scores are listed beside each network. Betweenness scores are not calculated for genes that do not connect to at least one other gene in the network (these genes are indicated with #N/A for betweenness). No ReactomeFI networks exist for the ‘organic anion, cation/zwitterion transport’, ‘metabolism of polyamines’, and ‘collagen formation’ pathways.
Crmps in Sema3A signaling

Degradation of the extracellular matrix

Sulfur amino acid metabolism

Lipid digestion, mobilization and transport
Signaling by PDGF

Signaling by TGF-β receptor complex

Gβγ signaling through PI3Kγ

Heparan sulfate heparin HS GAG metabolism
Signaling by Notch

Notch1 intracellular domain regulates transcription

Notch HLH transcription pathway
Signaling by Notch 1
PPARA activates gene expression

Triggering of complement

HDL mediated lipid transport

Lipoprotein metabolism
SMAD2_SMAD3_SMAD4 heterotrimer
Regulates transcription

PI3K_AKT activation

HS GAG biosynthesis
SMAD2_SMAD3_SMAD4 heterotrimer
Regulates transcription

HS GAG biosynthesis

PI3K_AKT activation

Nuclear receptor transcription pathway
Transcriptional activity of SMAD2_SMAD3_SMAD4 heterotrimer
<table>
<thead>
<tr>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 4</th>
<th>Level 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metabolism of lipids and lipoproteins</td>
<td>Lipoprotein metabolism</td>
<td>Lipoprotein metabolism</td>
<td>Oleic acid mediated lipid transport</td>
<td>Oleic acid mediated lipid transport</td>
</tr>
<tr>
<td>Metabolism of amino acids and derivatives</td>
<td>Sulfur amino acid metabolism</td>
<td>Metabolism of pyrimidines</td>
<td>Sulfur amino acid metabolism</td>
<td>Metabolism of purines</td>
</tr>
<tr>
<td>Metabolism of carbohydrates</td>
<td>Glycosaminoglycan metabolism</td>
<td>Glycosaminoglycan metabolism</td>
<td>N-acetyl neuraminic acid metabolism</td>
<td>N-acetyl neuraminic acid metabolism</td>
</tr>
<tr>
<td>Signaling by Notch</td>
<td>Notch signaling via Notch1 intracellular domain</td>
<td>Notch1 intracellular domain regulates transcription</td>
<td>Notch1 intracellular domain regulates transcription</td>
<td>Notch1 intracellular domain regulates transcription</td>
</tr>
<tr>
<td>Developmental Biology</td>
<td>Wnt signaling for neurite outgrowth</td>
<td>Wnt signaling for neurite outgrowth</td>
<td>Wnt signaling for neurite outgrowth</td>
<td>Wnt signaling for neurite outgrowth</td>
</tr>
<tr>
<td>Extracellular matrix organization</td>
<td>Collagen organization</td>
<td>Collagen organization</td>
<td>Collagen organization</td>
<td>Collagen organization</td>
</tr>
<tr>
<td>Immune system</td>
<td>Toll receptor cascades</td>
<td>Toll receptor cascades</td>
<td>Toll receptor cascades</td>
<td>Toll receptor cascades</td>
</tr>
<tr>
<td>Metabolism of proteins</td>
<td>Post-translational protein modification (PTM)</td>
<td>Post-translational protein modification (PTM)</td>
<td>Post-translational protein modification (PTM)</td>
<td>Post-translational protein modification (PTM)</td>
</tr>
<tr>
<td>Cell-cell communication</td>
<td>Cell-cell junction organization</td>
<td>Cell-cell junction organization</td>
<td>Cell-cell junction organization</td>
<td>Cell-cell junction organization</td>
</tr>
<tr>
<td>Transmembrane transport of small molecules</td>
<td>Neurotransmitter transport</td>
<td>Neurotransmitter transport</td>
<td>Neurotransmitter transport</td>
<td>Neurotransmitter transport</td>
</tr>
<tr>
<td>Gene Expression</td>
<td>Generic transcription pathway</td>
<td>Nuclear receptor transcription pathway</td>
<td>Nuclear receptor transcription pathway</td>
<td>Nuclear receptor transcription pathway</td>
</tr>
</tbody>
</table>
### Supplementary Table S2. Replication studies on 85 Discovery pathways using pruned lists of SNPs

SNPs were pruned if they were in LD, defined as $r^2 > 0.2$ or $r^2 > 0.5$. Pathways that also replicated with the pruned list are highlighted in yellow.

<table>
<thead>
<tr>
<th>Replicated pathways (SNPs with $r^2 &gt; 0.2$ pruned)</th>
<th>Number of genes</th>
<th>P</th>
<th>FDR</th>
<th>Replicated in unpruned list</th>
</tr>
</thead>
<tbody>
<tr>
<td>REACTOME_LIPID_DIGESTION_MOBILIZATION_AND_TRANSPORT</td>
<td>46</td>
<td>&lt; 0.001</td>
<td>0.007</td>
<td>Yes</td>
</tr>
<tr>
<td>REACTOME_MUSCLE_CONTRACTION</td>
<td>48</td>
<td>&lt; 0.001</td>
<td>0.013</td>
<td>No</td>
</tr>
<tr>
<td>REACTOME_EXTRACELLULAR_MATRIX_ORGANIZATION</td>
<td>87</td>
<td>&lt; 0.001</td>
<td>0.033</td>
<td>Yes</td>
</tr>
<tr>
<td>REACTOME_LIPOPROTEIN_METABOLISM</td>
<td>28</td>
<td>0.001</td>
<td>0.033</td>
<td>Yes</td>
</tr>
<tr>
<td>REACTOME_ORGANIC_CATION_ANION_ZWITTERION_TRANSPORT</td>
<td>13</td>
<td>0.002</td>
<td>0.012</td>
<td>No</td>
</tr>
<tr>
<td>REACTOME_SIGNALING_BY_PDGF</td>
<td>122</td>
<td>0.003</td>
<td>0.071</td>
<td>Yes</td>
</tr>
<tr>
<td>REACTOME_COLLAGENFORMATION</td>
<td>58</td>
<td>0.004</td>
<td>0.052</td>
<td>Yes</td>
</tr>
<tr>
<td>REACTOME_SMOOTH_MUSCLE_CONTRACTION</td>
<td>25</td>
<td>0.005</td>
<td>0.052</td>
<td>No</td>
</tr>
<tr>
<td>REACTOME_HDL_MEDIATED_LIPID_TRANSPORT</td>
<td>15</td>
<td>0.011</td>
<td>0.097</td>
<td>Yes</td>
</tr>
<tr>
<td>REACTOME_TOLL_RECEPTOR_CASCADES</td>
<td>118</td>
<td>0.012</td>
<td>0.105</td>
<td>Yes</td>
</tr>
<tr>
<td>REACTOME_SIGNALING_BY_RHO_GTPASES</td>
<td>113</td>
<td>0.017</td>
<td>0.134</td>
<td>No</td>
</tr>
<tr>
<td>REACTOME_PI_METABOLISM</td>
<td>48</td>
<td>0.022</td>
<td>0.107</td>
<td>No</td>
</tr>
<tr>
<td>REACTOME_NCAM_SIGNALING_FOR_NEURITE_OUT_GROWTH</td>
<td>64</td>
<td>0.022</td>
<td>0.157</td>
<td>Yes</td>
</tr>
<tr>
<td>REACTOME_DEGRADATION_OF_THE_EXTRACELLULAR_MATRIX</td>
<td>29</td>
<td>0.047</td>
<td>0.210</td>
<td>Yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Replicated pathways (SNPs with $r^2 &gt; 0.5$ pruned)</th>
<th>Number of genes</th>
<th>P</th>
<th>FDR</th>
<th>Replicated in unpruned list</th>
</tr>
</thead>
<tbody>
<tr>
<td>REACTOME_LIPID_DIGESTION_MOBILIZATION_AND_TRANSPORT</td>
<td>46</td>
<td>&lt; 0.001</td>
<td>0.001</td>
<td>Yes</td>
</tr>
<tr>
<td>REACTOME_LIPOPROTEIN_METABOLISM</td>
<td>28</td>
<td>&lt; 0.001</td>
<td>0.006</td>
<td>Yes</td>
</tr>
<tr>
<td>REACTOME_SULFUR_AMINO_ACID_METABOLISM</td>
<td>24</td>
<td>&lt; 0.001</td>
<td>0.010</td>
<td>Yes</td>
</tr>
<tr>
<td>REACTOME_CHYLOCOURN_MEDIATED_LIPID_TRANSPORT</td>
<td>16</td>
<td>&lt; 0.001</td>
<td>0.011</td>
<td>Yes</td>
</tr>
<tr>
<td>REACTOME_ORGANIC_CATION_ANION_ZWITTERION_TRANSPORT</td>
<td>13</td>
<td>0.002</td>
<td>0.009</td>
<td>No</td>
</tr>
<tr>
<td>REACTOME_DEGRADATION_OF_THE_EXTRACELLULAR_MATRIX</td>
<td>29</td>
<td>0.005</td>
<td>0.028</td>
<td>Yes</td>
</tr>
<tr>
<td>REACTOME_TRANSCRIPTIONAL_ACTIVITY_OF_SMAD2_SMAD3_SMAD4_HETEROTRIMER</td>
<td>38</td>
<td>0.005</td>
<td>0.068</td>
<td>Yes</td>
</tr>
<tr>
<td>REACTOME_NOTCH_HLH_TRANSCRIPTION_PATHWAY</td>
<td>13</td>
<td>0.011</td>
<td>0.099</td>
<td>Yes</td>
</tr>
<tr>
<td>REACTOME_SIGNALING_BY_TGF_BETA_RECEPTOR_COMPLEX</td>
<td>63</td>
<td>0.013</td>
<td>0.120</td>
<td>Yes</td>
</tr>
<tr>
<td>REACTOME_PI3K_AKT_ACTIVATION</td>
<td>38</td>
<td>0.02</td>
<td>0.126</td>
<td>Yes</td>
</tr>
<tr>
<td>REACTOME_TOLL_RECEPTOR_CASCADES</td>
<td>118</td>
<td>0.02</td>
<td>0.134</td>
<td>Yes</td>
</tr>
<tr>
<td>REACTOME_NOTCH1_INTRACELLULAR_DOMAIN_REGULATES_TRANSCRIPTION</td>
<td>46</td>
<td>0.02</td>
<td>0.140</td>
<td>Yes</td>
</tr>
<tr>
<td>REACTOME_SMAD2_SMAD3_SMAD4_HETEROTRIMER_REGULATES_TRANSCRIPTION</td>
<td>27</td>
<td>0.021</td>
<td>0.124</td>
<td>Yes</td>
</tr>
<tr>
<td>REACTOME_SYNTHESIS_OF_PI5S_AT_THE_PLASMA_MEMBRANE</td>
<td>31</td>
<td>0.023</td>
<td>0.145</td>
<td>No</td>
</tr>
<tr>
<td>Replicated Pathway Name</td>
<td># of SNPs</td>
<td>SNPs with pairwise $r^2 \geq 0.2$</td>
<td>SNPs with pairwise $r^2 \geq 0.5$</td>
<td>SNPs with pairwise $r^2 \geq 0.8$</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-----------</td>
<td>----------------------------------</td>
<td>----------------------------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>31 Cell Extracellular Matrix Interactions</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1 Chylocrion Mediated Lipid Transport</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>32 Collagen Formation</td>
<td>58</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2 CRMPs in Sem3A Signaling</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3 Degradation of the Extracellular Matrix</td>
<td>29</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15 Extracellular Matrix Organization</td>
<td>87</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>17 G Beta Gamma Signalling Through PI3KGamma</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>26 HDL Mediated Lipid Transport</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>21 Heparan Sulfate Heparin HS GAG Metabolism</td>
<td>52</td>
<td>2</td>
<td>2</td>
<td>2 (rs1509558, rs2663905)</td>
</tr>
<tr>
<td>13 HS GAG Biosynthesis</td>
<td>31</td>
<td>2</td>
<td>2</td>
<td>2 (rs1509558, rs2663905)</td>
</tr>
<tr>
<td>30 Initial Triggering of Complement</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4 Lipid Digestion Mobilization and Transport</td>
<td>46</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5 Lipoprotein Metabolism</td>
<td>28</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9 Metabolism of Polyamines</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20 NCAM Signaling for Neurite Out Growth</td>
<td>64</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>29 NCAM1 Interactions</td>
<td>39</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12 Notch HLH Transcription Pathway</td>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6 Notch1 Intracellular Domain Regulates Transcription</td>
<td>46</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>28 Nuclear Receptor Transcription Pathway</td>
<td>49</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11 Organic Cation Anion Zwitterion Transport</td>
<td>13</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>25 PI3K Akt Activation</td>
<td>38</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>24 PPARA Activates Gene Expression</td>
<td>104</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14 PTM Gamma Carboxylation Hypusine Formation and Arylsulfatase Activation</td>
<td>27</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>22 Signaling By Notch</td>
<td>103</td>
<td>2</td>
<td>2</td>
<td>2 (rs709308, rs12143347)</td>
</tr>
<tr>
<td>18 Signaling By Notch1</td>
<td>70</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>19 Signaling By PDGF</td>
<td>122</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>23 Signaling By TGF Beta Receptor Complex</td>
<td>63</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10 SMAD2 SMAD3 SMAD4 Heterotrimer Regulates Transcription</td>
<td>27</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7 Sulfur Amino Acid Metabolism</td>
<td>24</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8 TOLL Receptor Cascades</td>
<td>118</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16 Transcriptional Activity of Smad2 Smad3 Smad4 Heterotrimer</td>
<td>38</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>27 Transcriptional Regulation of White Adipocyte Differentiation</td>
<td>72</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cluster</td>
<td>Module Nodes in module</td>
<td>Node Percentage</td>
<td>Node (gene) list</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>------------------------</td>
<td>----------------</td>
<td>-----------------</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>144</td>
<td>0.2017</td>
<td>CTRB1, CTSG, ELANE, F10, F7, FURIN, GGCX, KLK2, KLKB1, MMP1, MMP10, MMP11, MMP13, MMP14, MMP15, MMP16, MMP17, MMP2, MMP24, MMP25, MMP3, MMP7, MMP8, MMP9, PLG, PROC, PROS1, PROZ, PRSS1, TIMP2, TPSAB1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>142</td>
<td>0.1989</td>
<td>TOR, RPS6KA1, RPS6KA2, RPS6KB2, SAA1, SOS1, SPTA1, SPTAN1, SPTB, SPTBN1, SPTBN2, SPTBN4, SPTBN5, SRC, STAT1, THEM4, TIAM2, TRIB3, TSC2, YWHAB</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>109</td>
<td>0.1527</td>
<td>RORB, RORC, RXRA, RXRB, RXRG, SLC27A1, SMARCD3, SREBF1, SREBF2, TBL1XR1, TEAD1, TEAD2, TEAD3, TEAD4, TGFβ1, TGS1, THRA, THRB, VDR, WWTR1, YAP1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>83</td>
<td>0.1162</td>
<td>, NRTN, P4HB, PARVA, PARVB, PCOLCE2, PDGFB, PLOD1, PLOD2, PLOD3, PPIB, PRNP, PSPN, PTK2, RSU1, SERPINH1, SPP1, ST8SIA4, TESK1, TLL1, TLL2, VASP</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>54</td>
<td>0.0756</td>
<td>IK3R4, PPM1A, RIPK1, RIPK2, SIGIRR, TBK1, TICAM1, TICAM2, TIRAP, TLR1, TLR10, TLR2, TLR3, TLR4, TLR5, TLR6, TLR9, TRAF3, TRAF6, UBE2N, ZFYVE20</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>43</td>
<td>0.0602</td>
<td>ADAM10, ADAM17, APH1A, APH1B, APP, ATP2A1, ATP2A2, ATP2A3, B4GALT1, CNTN1, DLL1, DLL4, DNER, DTX1, DTX2, DTX4, JAG1, JAG2, LFNG, MFNG, MIB1, NCSTN, NEURL, NOTCH2, NOTCH3, NOTCH4, NUMB, POFUT1, PSEN1, PSEN2, PSENEN, RAB6A, RFNG, SEL1L, ST3GAL3, ST3GAL6, TMED2, TNFRSF21</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>38</td>
<td>0.0532</td>
<td>A2M, CMA1, CTRB1, CTRB2, CTSQ, ELANE, F10, F7, FURIN, GGCX, KLK2, KLKB1, MMP1, MMP10, MMP11, MMP13, MMP14, MMP15, MMP16, MMP17, MMP2, MMP24, MMP25, MMP3, MMP7, MMP8, MMP9, PLG, PROC, PROS1, PROZ, PRSS1, TIMP2, TPSAB1</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>34</td>
<td>0.0476</td>
<td>ABCA1, ALB, AMN, APOA1, APOA2, APOA5, APOB, APOC2, APOE, CETP, CUBN, LDLR, LDLRAP1, LGMN, LIPC, LPA, LPL, MTTP, PDE1A, PLTP, SAR1B, SCARB1</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>22</td>
<td>0.0308</td>
<td>CDK5, CDK5R1, CRMP1, DPYSL2, DPYSL3, DPYSL4, DPYSL5, FES, NRP1, PLXNA1, PLXNA2, SEMA3A</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>12</td>
<td>0.0168</td>
<td>CDK5, CDK5R1, CRMP1, DPYSL2, DPYSL3, DPYSL4, DPYSL5, FES, NRP1, PLXNA1, PLXNA2, SEMA3A</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>0.0126</td>
<td>ARSA, ARSB, ARSG, ARSI, ARSJ, ARSK, SGSH, SUMF1, SUMF2</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>8</td>
<td>0.0112</td>
<td>EIF2C1, EIF2C2, EIF2C3, EIF2C4, MOV10, TNRC6A, TNRC6B, TNRC6C</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>6</td>
<td>0.0084</td>
<td>C1QA, C1QB, C1QC, C1S, C2, CRP</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>3</td>
<td>0.0042</td>
<td>ABCG1, ABCG5, ABCG8</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>0.0042</td>
<td>ETHE1, SQDRL, TST</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>0.0028</td>
<td>MRI1, MTAP</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>0.0028</td>
<td>CLPS, PNLIP</td>
<td></td>
</tr>
<tr>
<td>Pathway Description</td>
<td>P-value</td>
<td>FDR</td>
<td>Nodes (Genes)</td>
<td></td>
</tr>
<tr>
<td>----------------------------------------------------------------------------------</td>
<td>--------</td>
<td>------</td>
<td>------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>toll-like receptor 4 signaling pathway</td>
<td>0.0071</td>
<td></td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>toll-like receptor 2 signaling pathway</td>
<td>0.0066</td>
<td></td>
<td>67</td>
<td></td>
</tr>
<tr>
<td>transmembrane transport</td>
<td>0.024</td>
<td></td>
<td>244</td>
<td></td>
</tr>
<tr>
<td>innate immune response</td>
<td>0.0406</td>
<td></td>
<td>412</td>
<td></td>
</tr>
<tr>
<td>complement activation, lectin pathway</td>
<td>0.0006</td>
<td></td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>SCF-dependent proteasomal ubiquitin-dependent protein catabolic process</td>
<td>0.0014</td>
<td></td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>S-adenosylmethionine biosynthetic process</td>
<td>0.0003</td>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>activation of MAPK activity</td>
<td>0.0086</td>
<td></td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>regulation of Golgi inheritance</td>
<td>0.0003</td>
<td></td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Ras protein signal transduction</td>
<td>0.0065</td>
<td></td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>cell-matrix adhesion</td>
<td>0.0065</td>
<td></td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>dichotomous subdivision of terminal units involved in salivary gland branching</td>
<td>0.0005</td>
<td></td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>membrane protein ectodomain proteolysis</td>
<td>0.0012</td>
<td></td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>cell fate determination</td>
<td>0.0011</td>
<td></td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>response to exogenous dsRNA</td>
<td>0.0011</td>
<td></td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>phosphatidylinositol-3-phosphate biosynthetic process</td>
<td>0.0011</td>
<td></td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>negative regulation of sequence-specific DNA binding transcription factor activity</td>
<td>0.0053</td>
<td></td>
<td>54</td>
<td></td>
</tr>
<tr>
<td>negative regulation of cyclin-dependent protein kinase activity</td>
<td>0.0014</td>
<td></td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>nerve growth factor receptor signaling pathway</td>
<td>0.0194</td>
<td></td>
<td>197</td>
<td></td>
</tr>
<tr>
<td>axon guidance</td>
<td>0.0287</td>
<td></td>
<td>291</td>
<td></td>
</tr>
<tr>
<td>response to exogenous dsRNA</td>
<td>0.0011</td>
<td></td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>phosphatidylinositol-3-phosphate biosynthetic process</td>
<td>0.0011</td>
<td></td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>lung development</td>
<td>0.0061</td>
<td></td>
<td>62</td>
<td></td>
</tr>
<tr>
<td>negative regulation of NF-kappaB transcription factor activity</td>
<td>0.0046</td>
<td></td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>hippo signaling cascade</td>
<td>0.0024</td>
<td></td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>adenylate cyclase-inhibiting G-protein coupled receptor signaling pathway</td>
<td>0.0031</td>
<td></td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>axon guidance</td>
<td>0.0287</td>
<td></td>
<td>291</td>
<td></td>
</tr>
<tr>
<td>heparin biosynthetic process</td>
<td>0.0007</td>
<td></td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>cell fate determination</td>
<td>0.0011</td>
<td></td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>triglyceride metabolic process</td>
<td>0.0026</td>
<td></td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>membrane protein ectodomain proteolysis</td>
<td>0.0012</td>
<td></td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>cell fate determination</td>
<td>0.0011</td>
<td></td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>axon guidance</td>
<td>0.0287</td>
<td></td>
<td>291</td>
<td></td>
</tr>
<tr>
<td>lung development</td>
<td>0.0061</td>
<td></td>
<td>62</td>
<td></td>
</tr>
<tr>
<td>negative regulation of NF-kappaB transcription factor activity</td>
<td>0.0046</td>
<td></td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>cellular lipid metabolic process</td>
<td>0.011</td>
<td></td>
<td>112</td>
<td></td>
</tr>
<tr>
<td>inactivation of MAPK activity</td>
<td>0.0023</td>
<td></td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>transcription initiation from RNA polymerase II promoter</td>
<td>0.0139</td>
<td></td>
<td>141</td>
<td></td>
</tr>
<tr>
<td>androgen receptor signaling pathway</td>
<td>0.004</td>
<td></td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>adenylate cyclase-inhibiting G-protein coupled receptor signaling pathway</td>
<td>0.0031</td>
<td></td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>axon guidance</td>
<td>0.0287</td>
<td></td>
<td>291</td>
<td></td>
</tr>
<tr>
<td>lung development</td>
<td>0.0061</td>
<td></td>
<td>62</td>
<td></td>
</tr>
<tr>
<td>negative regulation of NF-kappaB transcription factor activity</td>
<td>0.0046</td>
<td></td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>cellular lipid metabolic process</td>
<td>0.011</td>
<td></td>
<td>112</td>
<td></td>
</tr>
<tr>
<td>inactivation of MAPK activity</td>
<td>0.0023</td>
<td></td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>transcription initiation from RNA polymerase II promoter</td>
<td>0.0139</td>
<td></td>
<td>141</td>
<td></td>
</tr>
<tr>
<td>Gene</td>
<td>FC</td>
<td>P-value</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>------</td>
<td>-----------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs3813505</td>
<td>2.42</td>
<td>0.0037</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1192789</td>
<td>2.60</td>
<td>0.0025</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs2071054</td>
<td>2.19</td>
<td>0.0065</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs7169463</td>
<td>1.66</td>
<td>0.0219</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs6795235</td>
<td>1.25</td>
<td>0.0561</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1751484</td>
<td>1.12</td>
<td>0.0764</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs2272670</td>
<td>0.93</td>
<td>0.1181</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1122540</td>
<td>1.90</td>
<td>0.0642</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs6955512</td>
<td>2.59</td>
<td>0.0026</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs3736868</td>
<td>3.17</td>
<td>0.0007</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs9589183</td>
<td>0.41</td>
<td>0.3866</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1283628</td>
<td>0.63</td>
<td>0.2329</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1749874</td>
<td>2.07</td>
<td>0.0084</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs2702068</td>
<td>4.17</td>
<td>0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs6737227</td>
<td>0.63</td>
<td>0.2333</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs7925220</td>
<td>1.75</td>
<td>0.0177</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1763944</td>
<td>1.95</td>
<td>0.0111</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1283628</td>
<td>0.63</td>
<td>0.2329</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1749874</td>
<td>2.07</td>
<td>0.0084</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs2702068</td>
<td>4.17</td>
<td>0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs6737227</td>
<td>0.63</td>
<td>0.2333</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs7925220</td>
<td>1.75</td>
<td>0.0177</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1763944</td>
<td>1.95</td>
<td>0.0111</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs2737635</td>
<td>13.943192</td>
<td>1.14E-14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>----------</td>
<td>----------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs7951382</td>
<td>rs2930894</td>
<td>rs7305929</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.5470014</td>
<td>0.000283791</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs8085335</td>
<td>0.8569665</td>
<td>0.139005991</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1696011</td>
<td>rs6795235</td>
<td>rs4346144</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.77813196</td>
<td>0.166674064</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs2358617</td>
<td>0.8980388</td>
<td>0.12646233</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs9404942</td>
<td>1.3144311</td>
<td>0.048480703</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.251648</td>
<td>0.056021151</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.74627846</td>
<td>0.179358315</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.8349746</td>
<td>0.014622625</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1214248</td>
<td>2.2568824</td>
<td>0.005534998</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1782142</td>
<td>rs2071054</td>
<td>1.1001407</td>
<td>0.079407095</td>
<td></td>
</tr>
<tr>
<td>1.2332739</td>
<td>0.058442141</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.6137883</td>
<td>0.24333899</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.4811152</td>
<td>0.033028194</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs3011220</td>
<td>1.0311185</td>
<td>0.093085375</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.9448202</td>
<td>0.001135481</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1158487</td>
<td>0.4916914</td>
<td>0.322335832</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1294517</td>
<td>0.81690025</td>
<td>0.152440289</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.054888</td>
<td>0.0000881</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1090882</td>
<td>1.6278546</td>
<td>0.023558377</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0479556</td>
<td>0.089545626</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1210602</td>
<td>rs2267958</td>
<td>1.9105166</td>
<td>0.012288061</td>
<td></td>
</tr>
<tr>
<td>rs1210436</td>
<td>1.6529106</td>
<td>0.022237677</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1192789</td>
<td>1.6904457</td>
<td>0.02039644</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.12951452</td>
<td>0.742139395</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs2306580</td>
<td>1.1957991</td>
<td>0.06370901</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1729165</td>
<td>rs1402696</td>
<td>rs6507876</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1108574</td>
<td>rs4802905</td>
<td>rs1781879</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.31812602</td>
<td>0.480699831</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1326848</td>
<td>rs1564183</td>
<td>rs3217738</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.1323864</td>
<td>0.000737248</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1738858</td>
<td>rs3804568</td>
<td>0.4916914</td>
<td>0.322335832</td>
<td></td>
</tr>
<tr>
<td>rs1472304</td>
<td>rs4759334</td>
<td>rs7618014</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1170485</td>
<td>rs7987737</td>
<td>0.92773265</td>
<td>0.118104745</td>
<td></td>
</tr>
<tr>
<td>rs1220754</td>
<td>0.57739145</td>
<td>0.2646114</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1527366</td>
<td>0.83786774</td>
<td>0.145255394</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.3686662</td>
<td>0.004278917</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1287612</td>
<td>rs1694680</td>
<td>rs4261877</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.6900789</td>
<td>0.020413671</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs2282544</td>
<td>rs1472304</td>
<td>rs1097140</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1190815</td>
<td>rs1317686</td>
<td>rs7031302</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs1214248</td>
<td>2.0478225</td>
<td>0.008957308</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2.7383418</td>
<td>0.001826662</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0036266</td>
<td>0.099168414</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs6073285</td>
<td>2.4280465</td>
<td>0.003732102</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.9735142</td>
<td>0.106288389</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.0198536</td>
<td>0.000955315</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs6094048</td>
<td>rs1171369</td>
<td>1.2468667</td>
<td>0.056641315</td>
<td></td>
</tr>
<tr>
<td>rs1188424</td>
<td>rs9404942</td>
<td>1.3870821</td>
<td>0.041012654</td>
<td></td>
</tr>
<tr>
<td>1.577712</td>
<td>0.026441614</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.1710087</td>
<td>0.067451444</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
MATERIALS AND METHODS

Study Population
We used the SNP level association p-value results from all 14 of the discovery GWAS analyzed by the transatlantic Coronary Artery Disease Genome wide Replication and Meta-analysis (CARDioGRAM) consortium as well as two GWAS conducted by the Ottawa Heart Institute. A summary of the core phenotypic details for these GWAS is presented in Table 1. We divided the 16 GWAS into 2 sets: a Stage 1 Discovery Set and a Stage 2 Replication Set. Further details on each of the cohorts included are provided in the Methods section in the online-only Data Supplement.

Genotyping and Quality Control
Details on genotyping and quality control procedures for all GWAS datasets included in this analysis have been previously reported. All datasets were restricted to subjects of white/European descent either through self-report, principal components analysis, or a combination of the two. Prior to performing association testing, all CARDioGRAM genome-wide datasets were imputed up to ~2.5 millions SNPs using HapMap2 release 22 (build 36) white (CEU; Center d’Etude du Polymorphisme Humain) sample as the reference panel. The SNP call rate filter used on genotyped SNPs pre-imputation was >0.90 and for a majority >0.95. These SNPs were then used to impute approximately 2.5 million HapMap SNPs. Imputed SNPs were then excluded based on missing frequency in cases or controls >0.02 (Missing), minor allele frequency in cases (MAF cases) or controls (MAF controls) < 0.01, quality of the imputation (INFO) < 0.5, and deviation from Hardy-Weinberg equilibrium in controls (HWE) p < 0.0001. One further filter that was applied to the discovery GWAS used in this analysis was the removal of any imputed SNP with an overall SNP call rate of < 0.75. The call rate was applied to imputed SNPs by only counting an imputed call if the posterior probability of one of the three possible genotypes was >90%, i.e. imputed with a high level of certainty. We also removed any SNP that was present in 2 or less of the Stage 1 studies and any SNP that was present in 4 or less of the Stage 2 studies.

Stage 1 and Stage 2 GWAS Meta Analysis
We conducted the meta-analysis of the two sets of GWAS using an analytic approach that was very similar to that used previously by the CARDioGRAM consortium. We then proceeded with a fixed-effects inverse-variance-weighted meta-analysis together with a Q- and I-measure of homogeneity. Any SNPs that were significantly heterogeneous based on Q and I statistics (at p<0.001) were analyzed using random effects.
Gene Set Enrichment Analysis

Pathway information for gene set enrichment analysis was obtained from the Reactome gene sets available in the Molecular Signatures Database v3.1 (MSigDB).\(^5\)\(^6\) Although a wide choice of pathway databases exist, we selected the expert-authored and manually curated Reactome database due to its transparent structural hierarchy, high generalizability and internally consistent ‘reaction-based’ data model encompassing a wide variety of biological processes. A total of 639 Reactome pathways were utilized for pathway enrichment analysis, after removing 35 pathways with <10 or >200 gene members\(^7\)\(^9\).

For SNP-to-gene mapping, a SNP, $S_i$, was first mapped to gene, $G_j$ ($j=1,\ldots,N$) if $S_i$ was located within the primary transcript of the gene or a window of 100 kilobases on either end of the gene (additional details in Figure S1 in the online-only Data Supplement). Associations between pathways and CAD were examined through a gene-set enrichment analysis (GSEA) procedure\(^5\)\(^10\) via the iGSEA4GWAS tool (Improved Gene Set Enrichment Analysis for Genome-Wide Association Study at http://gsea4gwas.psych.ac.cn/inputPage.jsp).\(^11\) iGSEA4GWAS examines the enrichment of significantly associated variants within or near a priori defined gene sets by determining if a particular gene-set ranks higher than a randomly distributed set, based on a running-sum statistic on the ranked list of genes (ranked by association $p$-values or an equivalent statistic) (Figure 1) (additional details in Methods section of online-only Data Supplement). The ‘improvement’ in iGSEA4GWAS over traditional GSEA approaches is realized by focusing on gene sets with high proportions of significant genes instead of relying solely on the overall gene set significance that may sometimes originate from only a few genes. Pathways achieving a permutation-based nominal $p$-value of $\leq0.05$ (at $<25\%$ false discovery rate, FDR) in the Stage 1 discovery studies were taken forward for replication in a meta-analysis of the Stage 2 studies. Pathways were a priori defined as replicated if they also achieved a $p$-value of $<0.05$ (corresponding to FDR$<12.5\%$) in the Stage 2 set of GWAS. Additional analysis was also conducted to test for the effects of linkage disequilibrium (LD) patterns among SNPs and among pathway genes on GSEA results (Methods section of online-only Data Supplement).

Bioinformatic Analysis to Prioritize Genes in Replicated Pathways

After identifying the replicated pathways, we sought to recover higher level functional interactions between the pathways, as well as between genes within a pathway. To accomplish this, we mapped the
genes from the replicated pathways onto well-curated interaction networks, and assessing the networks for (i) the probability that such networks can arise by chance (ii) the presence of biologically relevant clustering of genes within the network, and (iii) the relative contributions of pathway genes on the topology of pathway networks.

(i) Statistical evaluation of networks: To statistically evaluate the degree to which networks derived from the query genes could arise at random, we first mapped a total of 770 candidate genes (derived from the GWAS replicated pathways) to the highly curated and high confidence InWeb protein-protein interaction network (PPI)\textsuperscript{12}. Next, we created random networks via 1000 rounds of within-degree, node-label permutation of the InWeb PPI network, and compared parameters of network connectivity (node-degree and edge number) in sub-networks arising from the candidate genes in the original vs. the random networks. This analysis was conducted in the Disease Association Protein-Protein Link Evaluator (DAPPLE) software environment\textsuperscript{13}.

(ii) Mapping of replicated pathway genes to functionally interacting networks: The 770 genes from the replicated pathways were next mapped to a functional interaction network obtained from Reactome (ReactomeFI, 2012 version), and visualized in Cytoscape (v 2.8.2) (Methods section of online-only Data Supplement). The mapped interactome network was subjected to spectral partition clustering to identify internal modular sub-structures\textsuperscript{14}. We subsequently tested the resulting sub-networks for overrepresentation of biological processes via Gene Ontology-Biological Process terms.

(iii) Analysis of network topologies: Lastly, we converted the replicated Reactome pathways into functional network modules and analyzed their network centrality properties, via the Centiscape tool\textsuperscript{15}. Assuming that the critical functions of a network are largely governed by central nodes which connect several different neighborhoods of the network, we assessed the relative importance of networks and their constituent nodes (genes) by the centrality measures of ‘degree’ and ‘betweenness’ (Methods section , online-only Data Supplement).\textsuperscript{16} Although other network centrality descriptors exist, ‘degree’ and ‘betweenness’ have been proposed as key correlates of gene and protein function in biological networks. Of them, ‘betweenness’ has been proposed to the more relevant metric when studying network dysfunction in disease (genes with high ‘degree’ are usually essential for life and therefore may not be investigatable in the context of disease)\textsuperscript{17-19} More specifically, recent data further show that genes with intermediate connectivity (betweenness) have the highest probability of harboring germ-line
disease mutations (compared to essential genes) and correlate with pleiotropy, and crosstalk between functional modules.

Reference List


