SUPPLEMENTAL MATERIAL

Supplemental Methods

Smooth Muscle Cell Culture and Gene Expression

Smooth muscle cell (SMC) gene expression dataset and donor characteristics have been described in detail elsewhere[59]. Briefly, we cultured aortic SMCs isolated from 6, 12, 64, and 69individuals with East Asian, African, Admixed American, and European ancestries in complete media (containing 5% FBS) until 90% confluence. We then switched to either serum-free media for 24 hours to mimic the quiescent state of SMCs or continued to culture in complete media to mimic the proliferative state of SMCs[65]. Total RNA was extracted using the RNeasy Micro Kit (Qiagen) and the RNase-free DNase Set. RNA integrity scores for all samples, as measured by the Agilent TapeStation, were greater than 9, indicating high-quality RNA preparations. Sequencing libraries were prepared with the Illumina TruSeq Stranded mRNA Library Prep Kit and were sequenced to ~100 million read depth with 150 bp paired-end reads at the Psomogen sequencing facility. We trimmed the reads with low average Phred scores (<20) using Trim Galore[66] and mapped the reads to the hg38 version of the human reference genome using the STAR Aligner[67]. We quantified gene expression by calculating the transcripts per million (TPM) for each gene using RNA-SeQC based on GENCODE v32 transcript annotations. In addition to protein-coding RNAs, we also measured the non-coding RNA since they have been shown to play significant roles in SMC biology[68]. We considered a gene as expressed if it had more than 6 read counts and 0.1 TPM in at least 20% of the samples. RNAseq data is available from GEO with the accession number GSE193817.

Gene set enrichment analysis

We performed Gene Set Enrichment Analysis (GSEA) on all 11,330 expressed genes shared between the quiescent and proliferative conditions using GSEA software version 4.1 and predefined gene sets from the Molecular Signatures Database version 7.5[17,69], using only the 11,330 input genes as the background. A gene set is a group of genes that shares pathways, functions, chromosomal location, or other features. For the present study, we used the C5 ontology gene sets including Gene Ontology (GO) Biological Process and Molecular Function sets. GSEA ranks all of the genes in the dataset based on mean value differences and calculates gene set significance using an enrichment score defined as the maximum distance from the middle of the ranked list. The enrichment score indicates whether the genes contained in a gene set are clustered towards the beginning or the end of the ranked list. We used a False Discovery Rate (FDR) of 0.05 to signify enrichment, but included all pathways with an FDR ≤ 0.25 in **Supplemental Table II** comparing different pathway enrichment techniques.

Differential gene expression and functional enrichment analysis

We included genes with > 6 reads in at least 80% of the samples for both conditions for differential expression analysis using DESeq2[16]. Genes were differentially expressed between proliferative and quiescent conditions when $P_{adj} < 1x10^{-3}$ and $\log_2$(fold-change)>0.5. To

characterize the functional consequences of gene expression changes associated with proliferative and quiescent conditions, we performed gene set enrichment analysis on differentially expressed genes using Gene Ontology (GO) Biological Process and Molecular Function gene sets[18,19] with the anRichment R package[70]. We used an FDR of 0.05 to signify enrichment, but included all pathways with an FDR ≤ 0.25 in **Supplemental Table II** comparing different pathway enrichment techniques.

Weighted Gene Co-expression Network Analysis

A gene module is a cluster of densely interconnected genes in terms of co-expression. We used Iterative Weighted Gene Co-expression Network Analysis (iterativeWGCNA)[10], which uses hierarchical clustering and an adjacency matrix, to identify gene modules. The adjacency matrix is defined as the similarity between the i-th gene and j-th gene based on the absolute value of the Pearson correlation coefficient between the profiles of genes i and j. Soft-thresholding powers are applied to the adjacency matrix in order to reduce the noise of correlations and create a network that resembles a scale-free graph representative of biological systems[12,13]. Scale-free graphs are characterized by a power-law distribution where few hub nodes exist and new nodes prefer to connect with existing nodes. Too low of a soft-threshold power may lead to high correlation among large groups of genes in a dataset invalidating the assumption of the scale-free topology approximation[14]. Conversely, if too high of a soft-threshold power is used, hub nodes that are present under the assumption of scale-free topology may lose connectivity and diminish biologically relevant co-expression networks. Therefore, following WGCNA[9] recommendations, we chose the lowest soft-thresholding power that generates a node connectivity distribution representative of scale-free topology[14]. IterativeWGCNA follows the same principles as WGCNA but re-runs WGCNA iteratively to prune poorly fitting genes resulting in more refined modules compared to WGCNA. Genes that are not assigned to any of the modules are designated to the grey module. Because these genes are not co-expressed, we did not consider them in our analyses.

Network preservation analysis

We performed preservation analysis on modules constructed using iterativeWGCNA to study their changes across the two cell culture conditions. To determine whether a pathway of genes is perturbed between the proliferative and quiescent conditions, we studied modules whose connectivity patterns are not preserved between conditions as demonstrated by their module preservation statistics. For this analysis we used the summary statistic, medianRank, implemented in the WGCNA R package as a composite module preservation statistic[15]. medianRank is a rank-based measure that relies on observed preservation statistics. medianRank is calculated as the mean of medianRank.density and medianRank.connectivity. Density is the mean adjacency (connection strength) across all nodes in the network. Connectivity is the sum of connection strengths with the other network nodes. To calculate medianRank.density and medianRank.connectivity, for each statistic d in the reference network, we ranked modules in

the test network based on the observed values obs_a^q. Thus, each module is assigned a rank $rank_a^q$ or each observed statistic. The median density and connectivity ranks are then calculated for each module, q, in the test network. The test and reference networks were then flipped to calculate preservation for each condition in the other. A module with a lower medianRank exhibits stronger observed preservation statistics than a module with a higher median rank. We identified the least preserved modules by defining the modules scoring in the bottom 20th percentile of preservation (modules with the highest medianRank score).

Coronary Artery Disease-associated gene sets

Genome-wide association studies (GWAS) alone are unable to identify the causal gene at a locus[71]; consequently, there are oftentimes several genes that are potentially causal at a given loci. Many fine-mapping and gene prioritization strategies have been created to try and determine causal genes, but the majority of these loci only have predicted causal genes. In order to capture all genes potentially involved in CAD, we used two different curated gene sets based on the 175 genomic loci associated with coronary artery disease (CAD) risk through GWAS[72]. The CAD Candidate gene set includes 2051 genes representing all genes in and near the 175 CAD GWAS loci. The CAD Prioritized gene set contains 175 genes predicted to be causal at each genome-wide significant loci based on functional annotation, such as genomic location, biological pathway interpretation, literature reviews, and DEPICT gene prioritization[73]. Our dataset included 956 genes from the CAD Candidate gene set and 104 genes from the CAD Prioritized gene set.

Pathway enrichment of co-expression modules

To interpret the biological significance of the co-expression modules in the top 20% percentile and the bottom 20% percentile of preservation, we performed enrichment analysis using GO Biological Process and Molecular Function gene sets with the anRichment R package. We used an FDR cutoff of 0.05 to signify enrichment, but included all pathways with an FDR ≤ 0.25 in **Supplemental Table II** comparing different pathway enrichment techniques.

Bayesian Network Construction

Bayesian networks are directed acyclic graphs in which the edges of the graph are defined by conditional probabilities that characterize the distribution of states of each node given the state of its parents[74]. The joint probability distribution $p(X)$ on a set of nodes X is represented by $p(X) = \prod_i p(X^i|Pa(X^i))$, where $Pa(X^i)$ represents the parent set of X^i. In reconstructing Bayesian networks of gene expression data, each node represents a quantitative trait which is the expression level of a gene. We used expression levels of genes identified in co-expression modules as input into the Reconstructing Integrative Molecular Bayesian Networks (RIMBANet) algorithm[33,75,76]. The RIMBANet shell script[77] was adapted for implementation on University of Virginia's computing cluster triggering the parameters below:

-d discretized gene expression data
-b RIMBANet Bayesian network program path
-o output directory for Bayesian network
-e cis eQTL data file; EMPTY
-C use continuous data to update prior; TRUE
-w continuous gene expression data

Continuous expression data was used for calculating partial priors, which are then used as priors in the network construction[78]. Since we cannot give prior probability to every acyclic digraph, the continuous data generates a connectivity matrix that assesses some degree of believe over the dependency between two variables[79]. Discretized gene expression data was used for Bayesian network reconstruction. The data was discretized into three states for each gene: high expression levels, medium expression levels, and low expression levels (including unexpressed). Discretization allows for both linear and non-linear interactions to be captured and is computationally more efficient than using continuous data[33].

For each Bayesian network we reconstructed 1,000 Bayesian networks starting with 1,000 different randomly generated seeds. Markov chain Monte Carlo simulations are employed to identify thousands of different plausible networks. Small random changes are made to each network by flipping, adding, or deleting individual edges, ultimately accepting those changes that lead to an overall improvement in the fit of the network to the data, represented by the Bayesian Information Criterion[80]. Edges that appeared in greater than 30% of the networks were used to define a consensus network. Edges that were involved in loops were then removed from the consensus network.

RIMBANet performs well within large datasets (Node > 50) resulting in high true positive rates and precision, but other BN reconstruction methods have been shown to perform better for networks with smaller sets of nodes[81]. For networks less than 50 nodes, we implemented the bnlearn[52] R package to validate network predictions from RIMBANet. We utilized a constraint-based algorithm, incremental association (IAMB), that learns the network structure by analyzing the probabilistic relations entailed by the Markov property of Bayesian networks using the same gene expression data from genes identified in co-expression modules used as input into RIMBANet. The IAMB algorithm was performed using an optimized implementation (default settings) that uses backtracking to roughly halve the number of independence tests.

Key Driver Analysis

To identify key regulators for a given regulatory network, we performed key driver analysis (KDA)[34], which takes as input a set of genes (G) and a directed gene network (N). KDA first generates a sub-network N_G, defined as the set of nodes in N that are no more than h-layers

away from the nodes in G. We first computed the size of the h-layer neighborhood (HLN) for each node in the reconstructed BN. For the given network N, μ was defined as the average size of the HLN. A score was added for a specific node if the HLN was greater than $\mu + \sigma(\mu)$. Total key driver scores for each node were then defined as the summation of all scores at each h-layer scaled according to h.

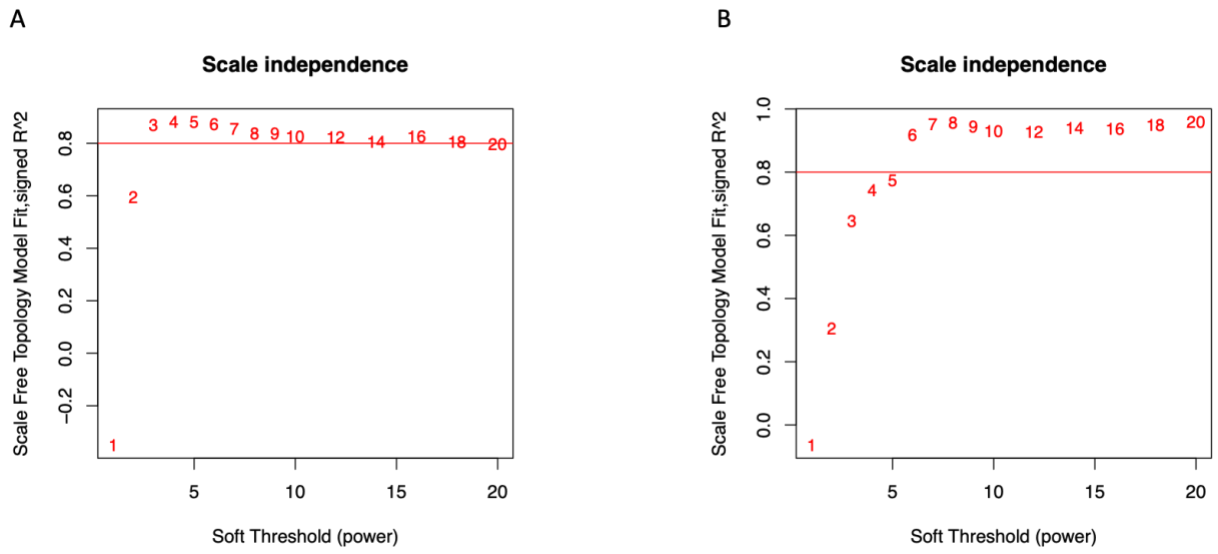
Pathway visualization of differentially expressed genes

We used the Pathview R package[82] to visualize genes differentially expressed between quiescent and proliferative VSMCs for the KEGG pathways of Nitrogen Metabolism and Glycolysis/Gluconeogenesis[43,44]. Genes present in both our RNAseq dataset and KEGG pathways are visualized as colored rectangular nodes. Green nodes represent downregulated genes and red nodes represent upregulated genes in the proliferative condition based off of log2foldchange between the two conditions. Grey nodes represent genes that were not differentially expressed.

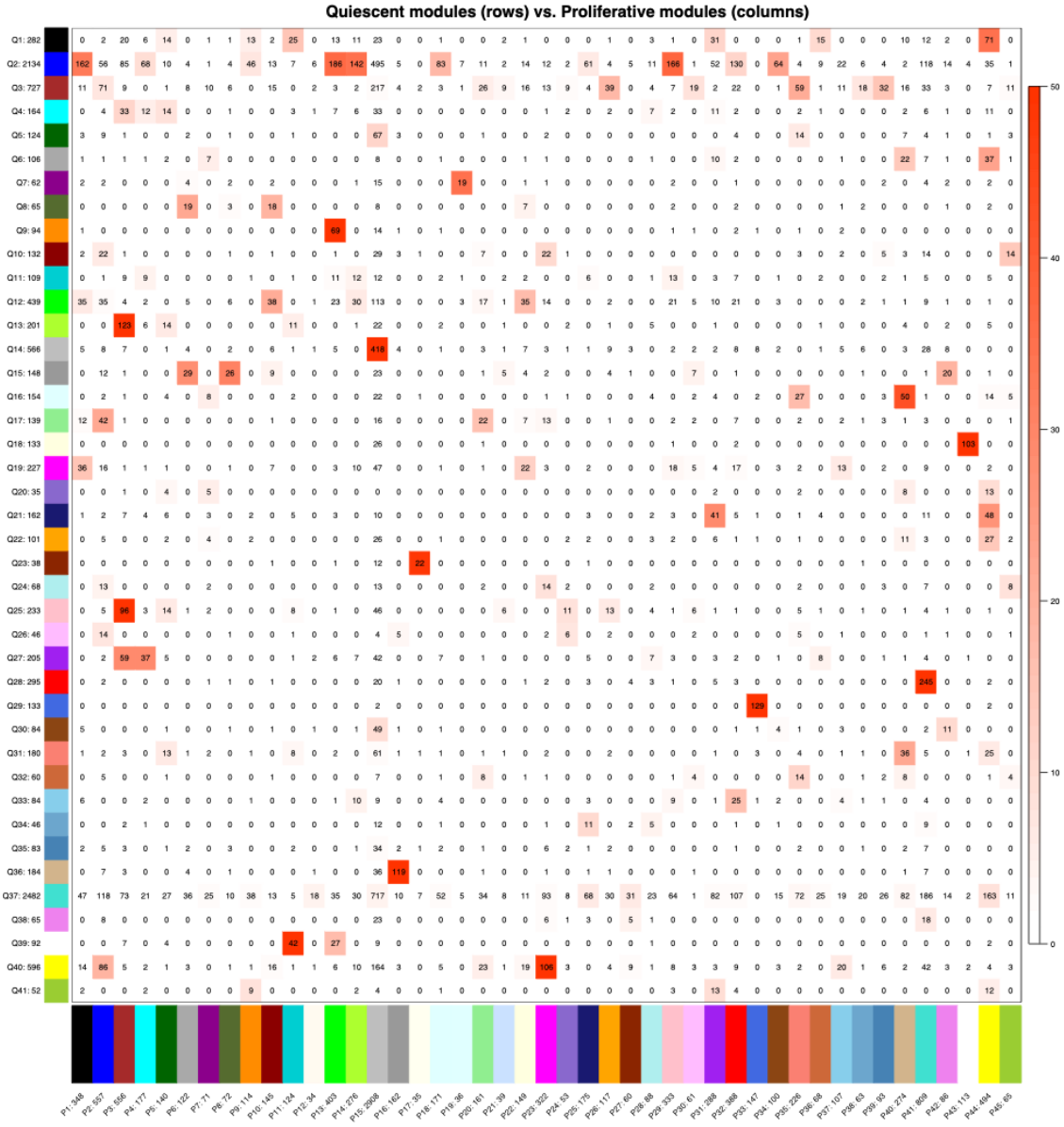
Hypergraph Models

We used the HyperG R package to create hypergraphs that visually represent the biological pathway differences between two BNs. First, we identified enriched GO terms (FDR ≤ 0.05) within each BN using PANTHER version 14[83]. In the hypergraphs, each of the associated GO terms are represented by a numbered node. The size of the node directly relates to the number of genes in the BN that are also a member of a specific GO term. Edges that encompass nodes represent the genes present in enriched GO terms. If multiple edges surround a node, all genes assigned to those edges are a part of the GO term. A floating node means there was no enrichment for the associated GO term in that BN. Edges and gene labels are color coordinated.

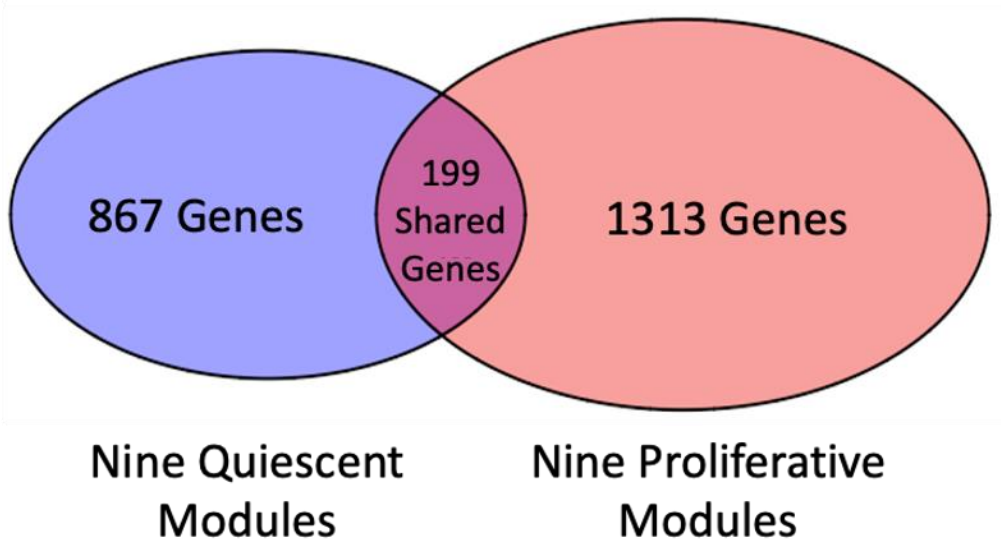
Supplemental Figures and Figure Legends



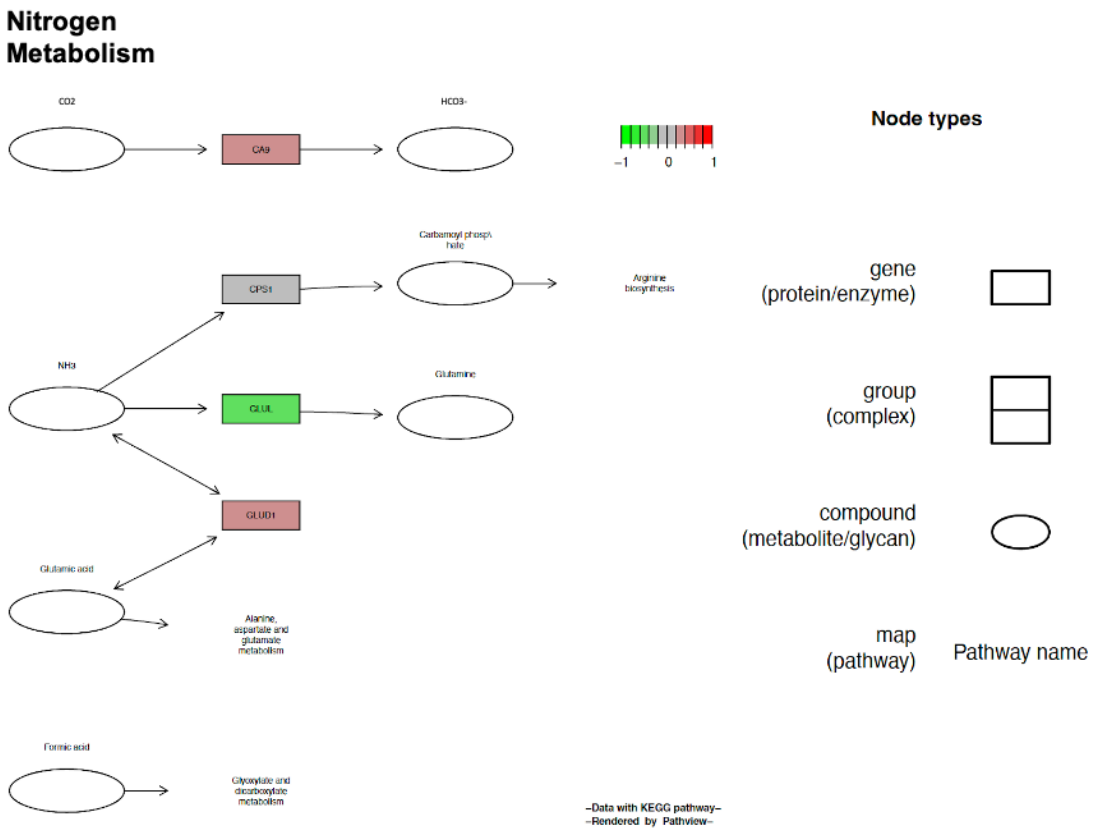
Supplemental Figure I: Determination of soft-thresholding power in Weighted Gene Co-expression Network Analysis. The scale-free fit index (y-axis) as a function of the soft thresholding power (x-axis) for gene expression of (A) quiescent and (B) proliferative smooth muscle cells. An R^2 value of 0.8 was used as the cutoff corresponding to scale-free topology.



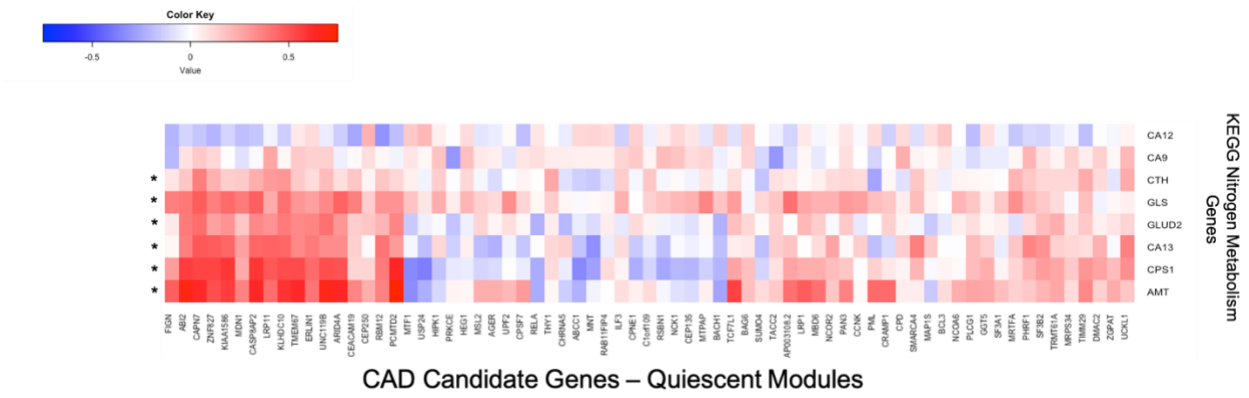
Supplemental Figure II: Overlap table of genes shared in modules across quiescent and proliferative conditions. Cross tabulation of quiescent modules (rows) and proliferative modules (columns). Each row and column is labeled by the corresponding module color and the total number of genes in the intersection of the corresponding row and column module. The table is color-coded by the Fisher exact test p-value of the overlap of gene module membership (-log(p)), according to the color legend on the right.



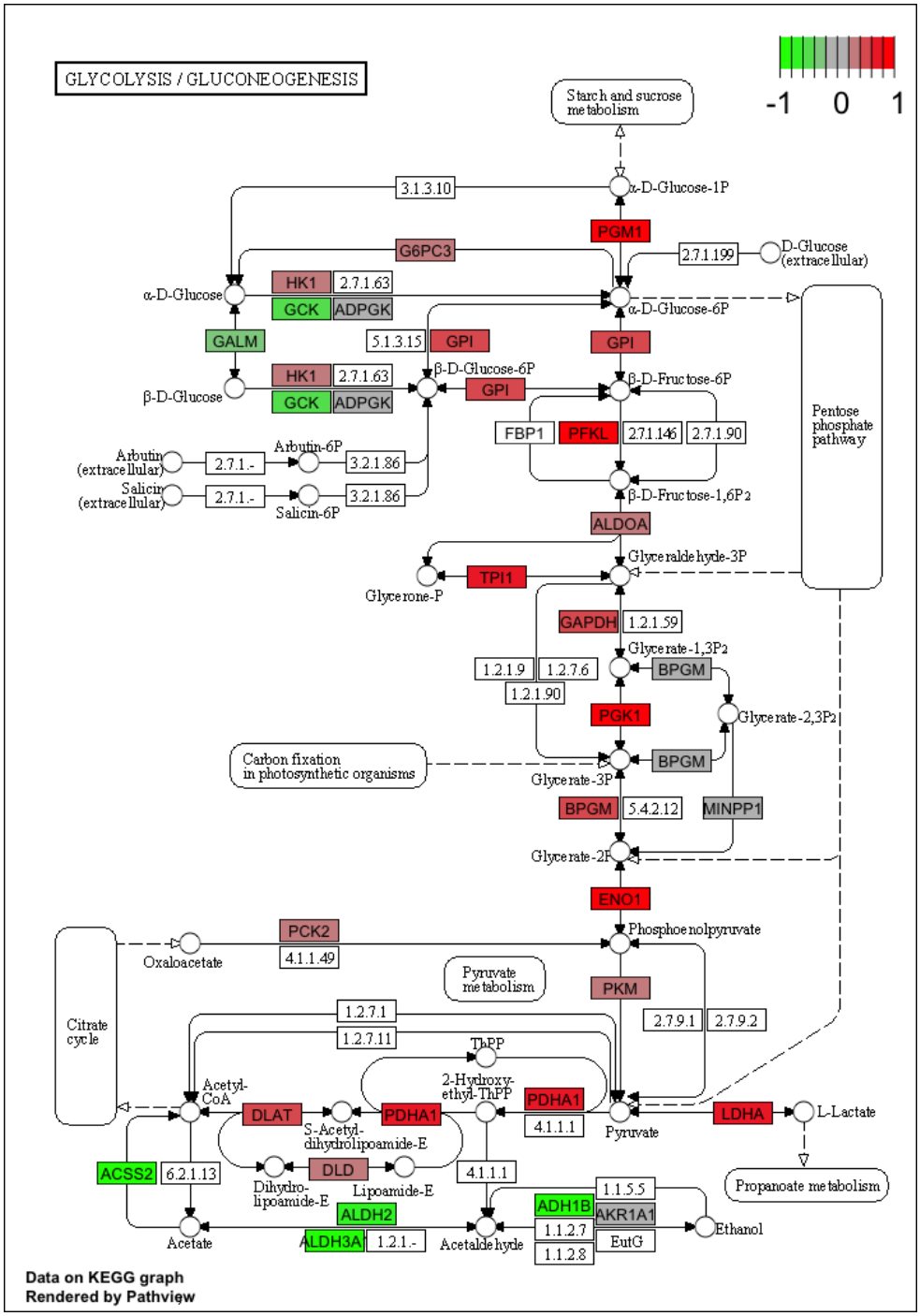
Supplemental Figure III: Representation of genes in the least preserved modules. A Venn diagram of the genes contained in the 18 least preserved modules.



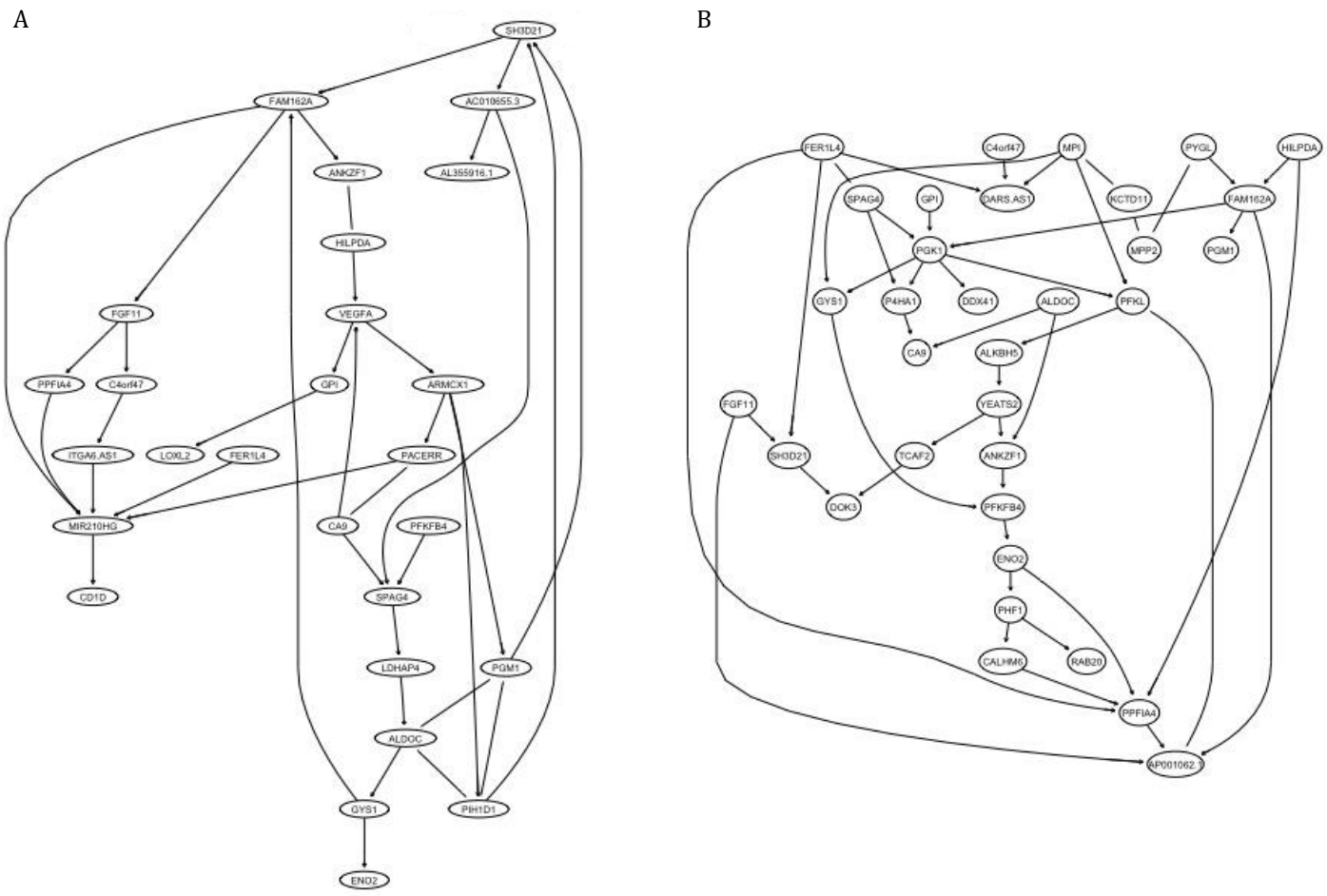
Supplemental Figure IV: Pathway visualization of differentially expressed genes between quiescent and proliferative VSMCs in the Nitrogen Metabolism KEGG pathway. Rectangular nodes represent protein coding genes. Red nodes signify genes upregulated in proliferative VSMCs, and green nodes signify genes downregulated in proliferative VSMCs. Grey nodes represent genes not differentially expressed.



Supplemental Figure V: Correlations between nitrogen-related genes and CAD-associated genes in quiescent VSMC gene expression. Heatmap of Pearson correlations (r) between 69 CAD candidate genes in the Q1, Q4, Q6, Q13, and Q38 modules and 8 genes in the KEGG Nitrogen Metabolism pathway. Asterisk marks denote genes in the KEGG Nitrogen Metabolism pathway with at least one correlation (r) $\geq |0.3|$ at a Bonferroni corrected p-value $\leq 5 \times 10^{-5}$.



Supplemental Figure VI: Pathway visualization of differentially expressed genes between quiescent and proliferative VSMCs in the Glycolysis/Gluconeogenesis KEGG pathway. Rectangular nodes represent protein coding genes. Red nodes signify genes upregulated in proliferative VSMCs, and green nodes signify genes downregulated in proliferative VSMCs. Grey nodes represent genes not differentially expressed.



Supplemental Figure VII: Bayesian networks of genes in the (A) Q23 module and (B) P17 module generated using the bnlearn R package.