Cell Reports, Volume *21*

Supplemental Information

Hepatic Dysfunction Caused by Consumption of a High-Fat Diet

Anthony R. Soltis, Norman J. Kennedy, Xiaofeng Xin, Feng Zhou, Scott B. Ficarro, Yoon Sing Yap, Bryan J. Matthews, Douglas A. Lauffenburger, Forest M. White, Jarrod A. Marto, Roger J. Davis, and Ernest Fraenkel

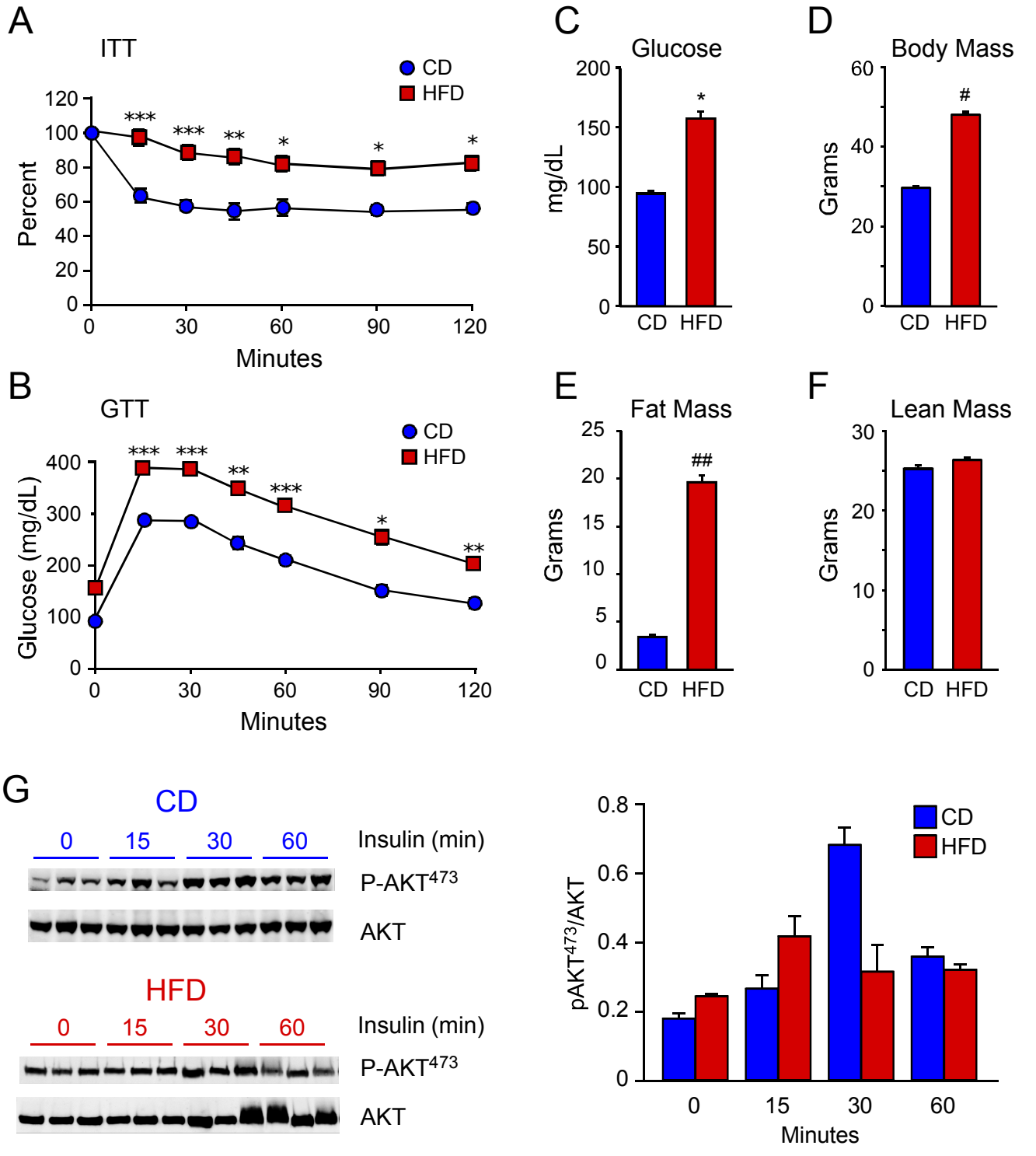

A
ITT
CD
HFD
Percent
Minutes
B
GTT
Glucose (mg/dL)
Minutes
C
Glucose
mg/dL
D
Body Mass
Grams
E
Fat Mass
Grams
F
Lean Mass
Grams
G
CD
Insulin (min)
P-AKT473
AKT
HFD
pAKT473/AKT
Minutes
H
H&E
Oil Red O
Lipid
Nuclei
100 μm

Figure S1

Figure S1. Physiological analysis of CD-fed and HFD-fed mice and analysis of hepatic steatosis (related to Figure 2).

(A, B) Insulin tolerance (ITT, *A*) and glucose tolerance (GTT, *B*) tests were performed using CD-fed and HFD-fed (16 weeks) mice (mean ± SEM; n = 20; *, $p < 1e\text{-}06$; **, $p < 5e\text{-}07$; ***, $p < 1e\text{-}07$).

(C) CD-fed and HFD-fed mice were fasted overnight and the blood concentration of glucose was measured (mean ± SEM; n=20; *, $p < 1e\text{-}06$).

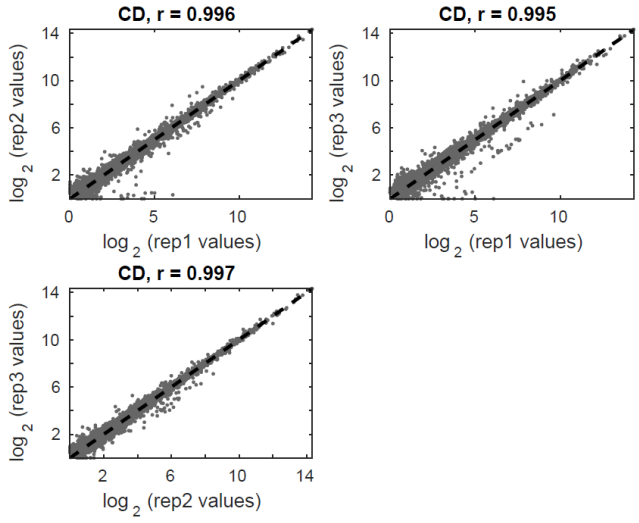
(D-F) The body mass of CD-fed and HFD-fed mice was measured (mean ± SEM; n=20; #, $p < 5e\text{-}13$). Fat and lean mass were measured by 1H-MRS analysis (mean ± SEM; n=20; ##, $p < 5e\text{-}16$).

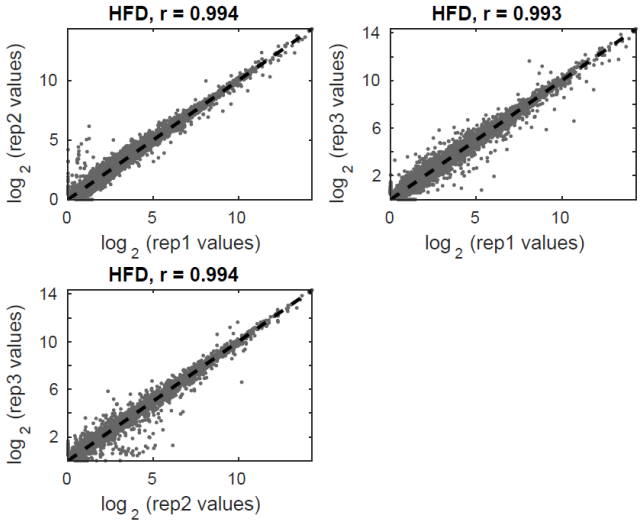
(G) CD-fed and HFD-fed mice were treated with insulin (1 U/kg) by intraperitoneal injection. Hepatic AKT was examined by immunoblot analysis by probing with antibodies to pSer473-AKT and AKT (mean ± SEM; n = 3). The data are normalized to total AKT.

(H) Consumption of a HFD causes hepatic steatosis: sections of the liver from CD-fed and HFD-fed mice were stained with hematoxylin & eosin (H&E) or with Oil Red O. The images are representative of sections prepared from three mice.

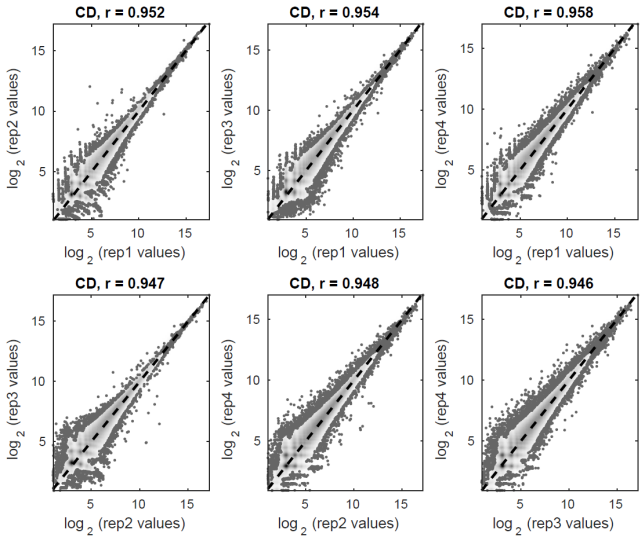
Figure S2

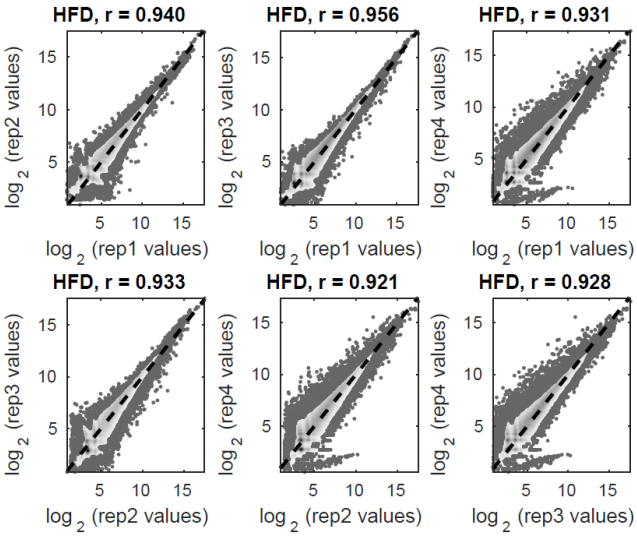
A - mRNA-Seq datasets



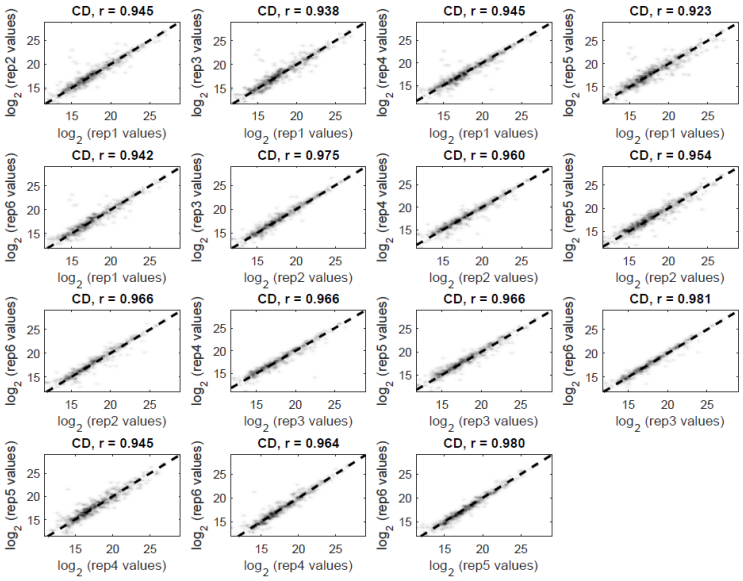


B - Global proteomic datasets





C - metabolomics datasets



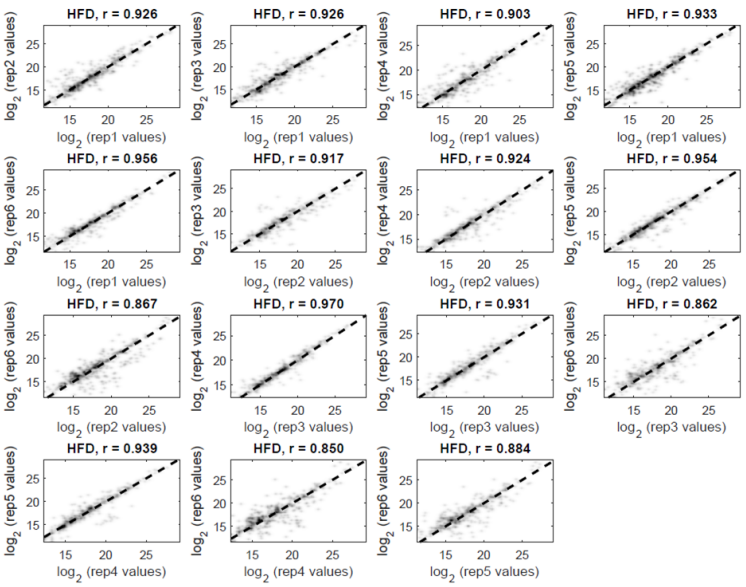


Figure S2. Replicate correlations for mRNA-Seq, global proteomic, and metabolomic datasets (related to Figure 2).

(A) Replicate correlations for mRNA-Seq samples (left – CD replicates, n = 3, $\log_2$ normalized read counts; right – HFD replicates, n = 3, $\log_2$ normalized read counts).

(B) Replicate correlations for global proteomic samples (left – CD replicates, n = 4, $\log_2$ normalized iTRAQ levels; right – HFD replicates, n = 4, $\log_2$ normalized iTRAQ levels).

(C) Replicate correlations for metabolomics samples (left – CD replicates, n = 6, $\log_2$ normalized abundances; right – HFD replicates, n = 6, $\log_2$ normalized abundances).

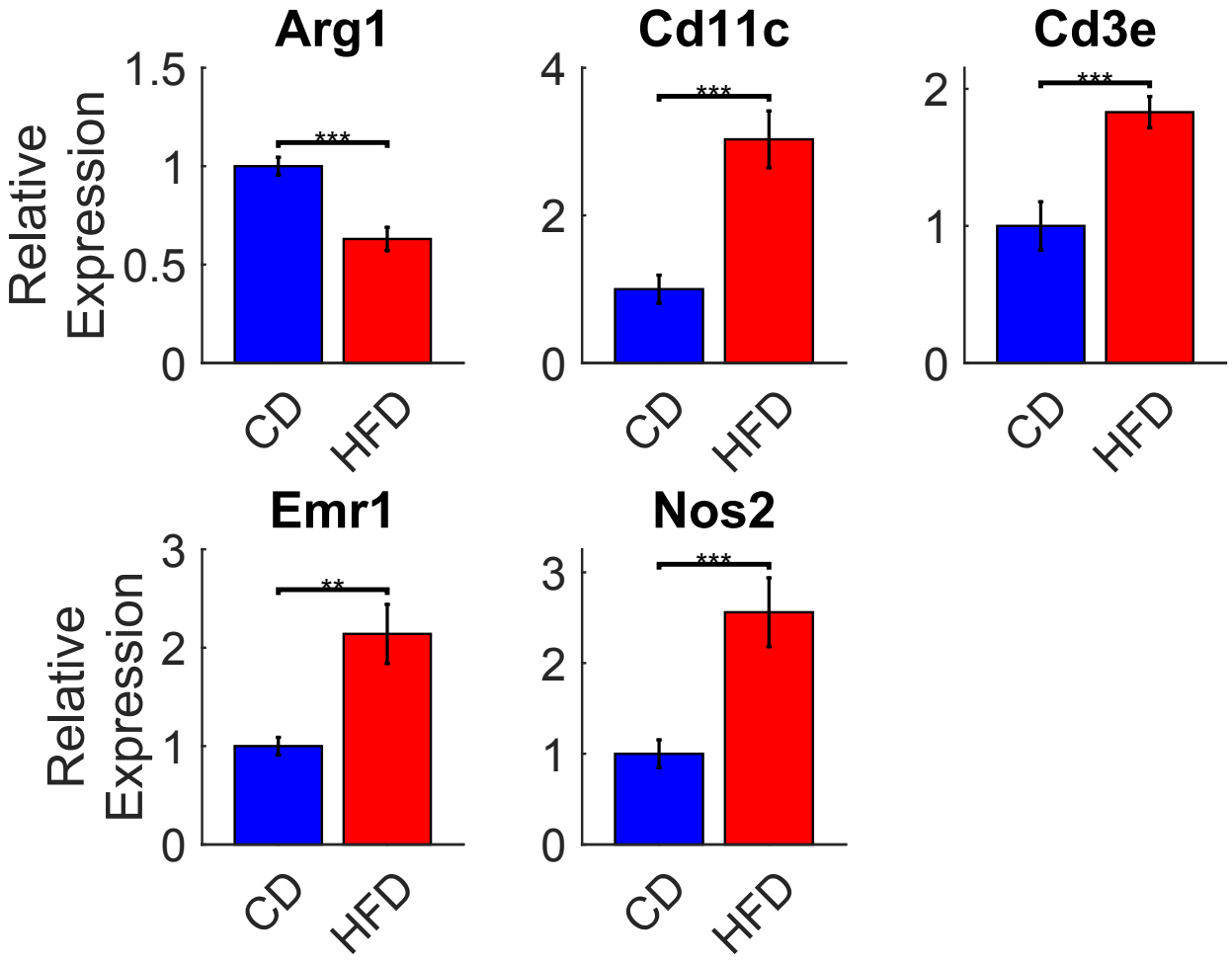
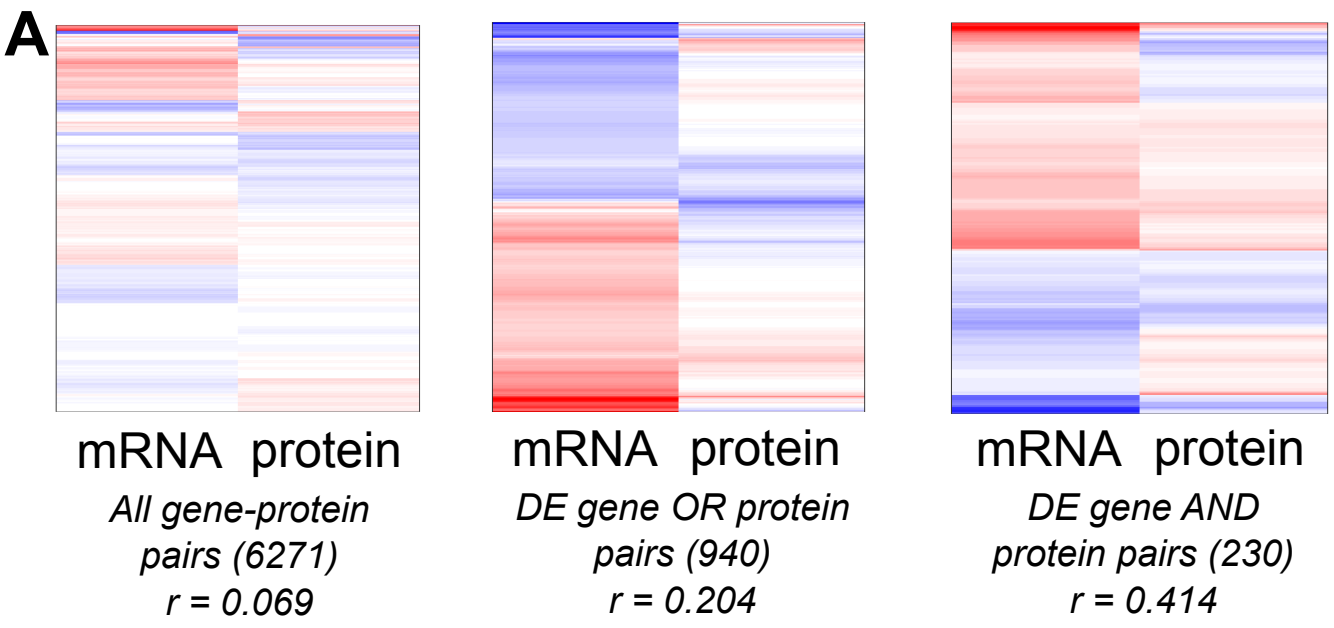
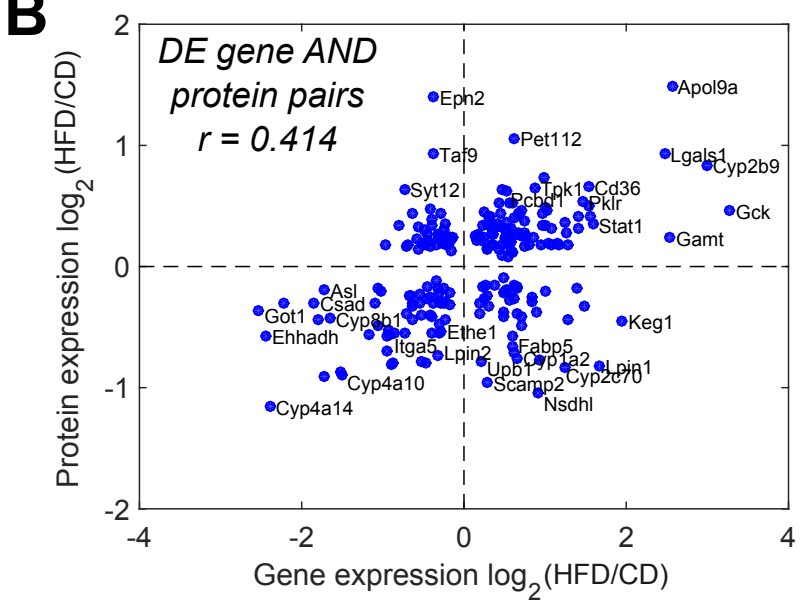


Figure S3. TaqMan assays for immune cell marker genes(related to Figure 2).

We performed TaqMan assays on immune cell marker genes to further establish evidence for the role of inflammatory processes in promoting and maintaining the insulin resistant state following HFD. All comparisons are between nine independent liver samples for each condition, except for *Emr1*which used eight livers per condition. *** $p < 0.001$, ** $p < 0.01$.

Figure S4


A
mRNA protein
All gene-protein pairs (6271)
r = 0.069
mRNA protein
DE gene OR protein pairs (940)
r = 0.204
mRNA protein
DE gene AND protein pairs (230)
r = 0.414
- down HFD
+ up HFD


B
DE gene AND protein pairs
r = 0.414
Protein expression log2 (HFD/CD)
Gene expression log2 (HFD/CD)
Epn2
Apol9a
Pet112
Taf9
Lgals1
Cyp2b9
Syt12
Tpk1
Cd36
Pcbd1
Pklr
Gck
Stat1
Gamt
Asl
Csad
Got1
Cyp8b1
Ehhadh
Ethe1
Itga5
Lpin2
Fabp5
Cyp1a2
Upb1
Scamp2
Cyp2c70
Lpin1
Keg1
Cyp4a10
Cyp4a14
Nsdhl

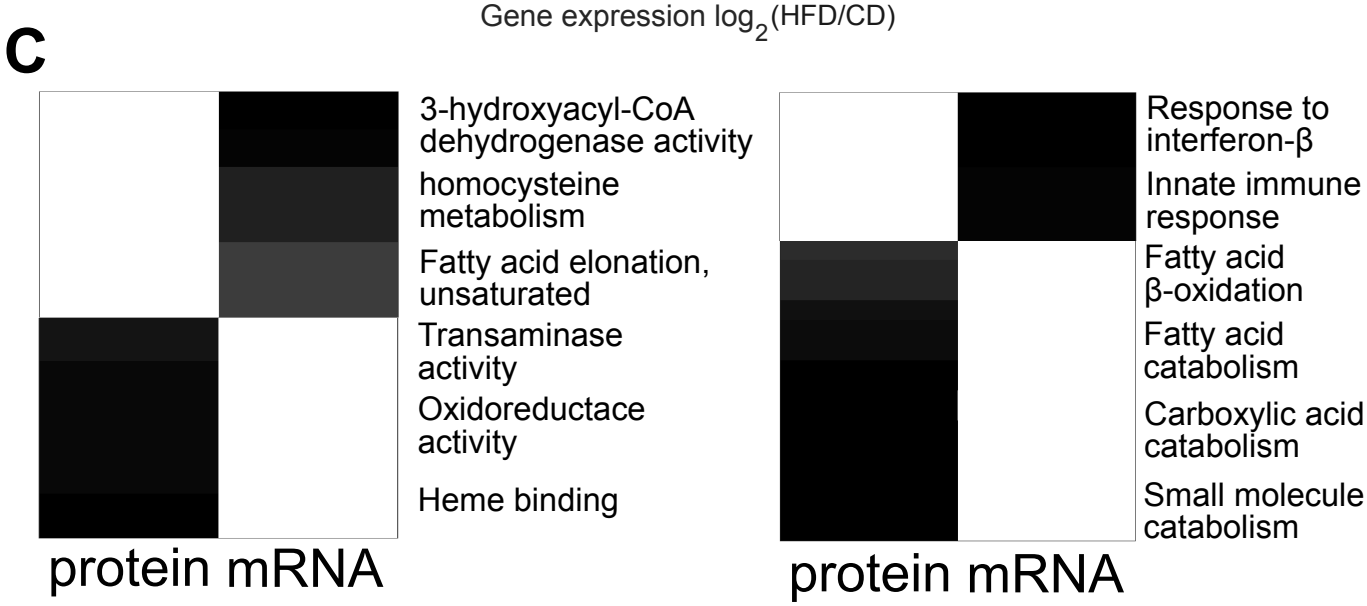

C
3-hydroxyacyl-CoA dehydrogenase activity
homocysteine metabolism
Fatty acid elonation, unsaturated
Transaminase activity
Oxidoreductace activity
Heme binding
protein mRNA
Down in HFD vs. CD
Gene Ontology
Response to interferon-β
Innate immune response
Fatty acid β-oxidation
Fatty acid catabolism
Carboxylic acid catabolism
Small molecule catabolism
protein mRNA
Up in HFD vs. CD
Gene Ontology

Figure S4. Comparison of gene and protein expression changes between CD and HFD livers (related to Figure 2).

(A) Clustergrams of mRNA and protein $\log_2$ fold-changes between CD and HFD livers restricted to all observed gene-protein pairs (left), differential genes or proteins (middle), and differential genes and proteins (right). Pearson correlation coefficients are shown for all comparisons.

(B) Scatter plot of same data at right in panel (A), highlighting individual species.

(C) Comparison of gene ontology enrichments specific to either mRNA or proteomic data in species that were either down-regulated (left) or up-regulated (right) by HFD.

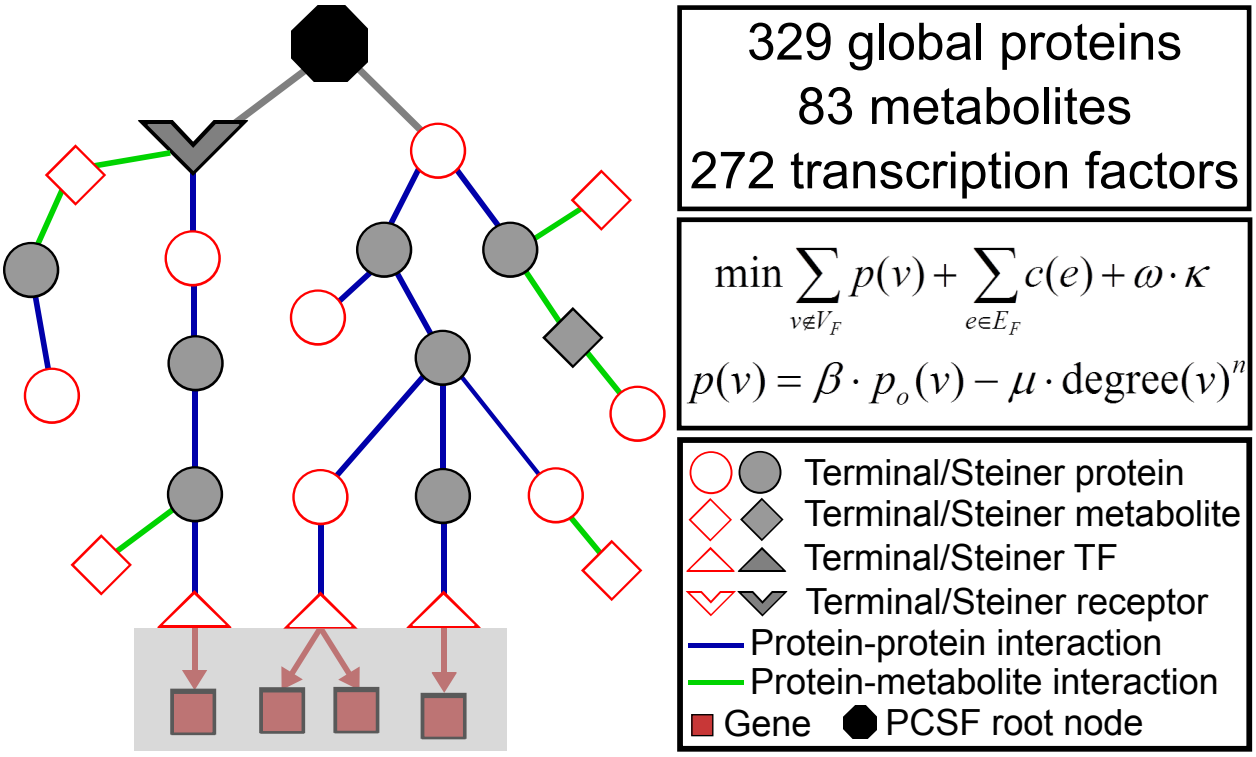
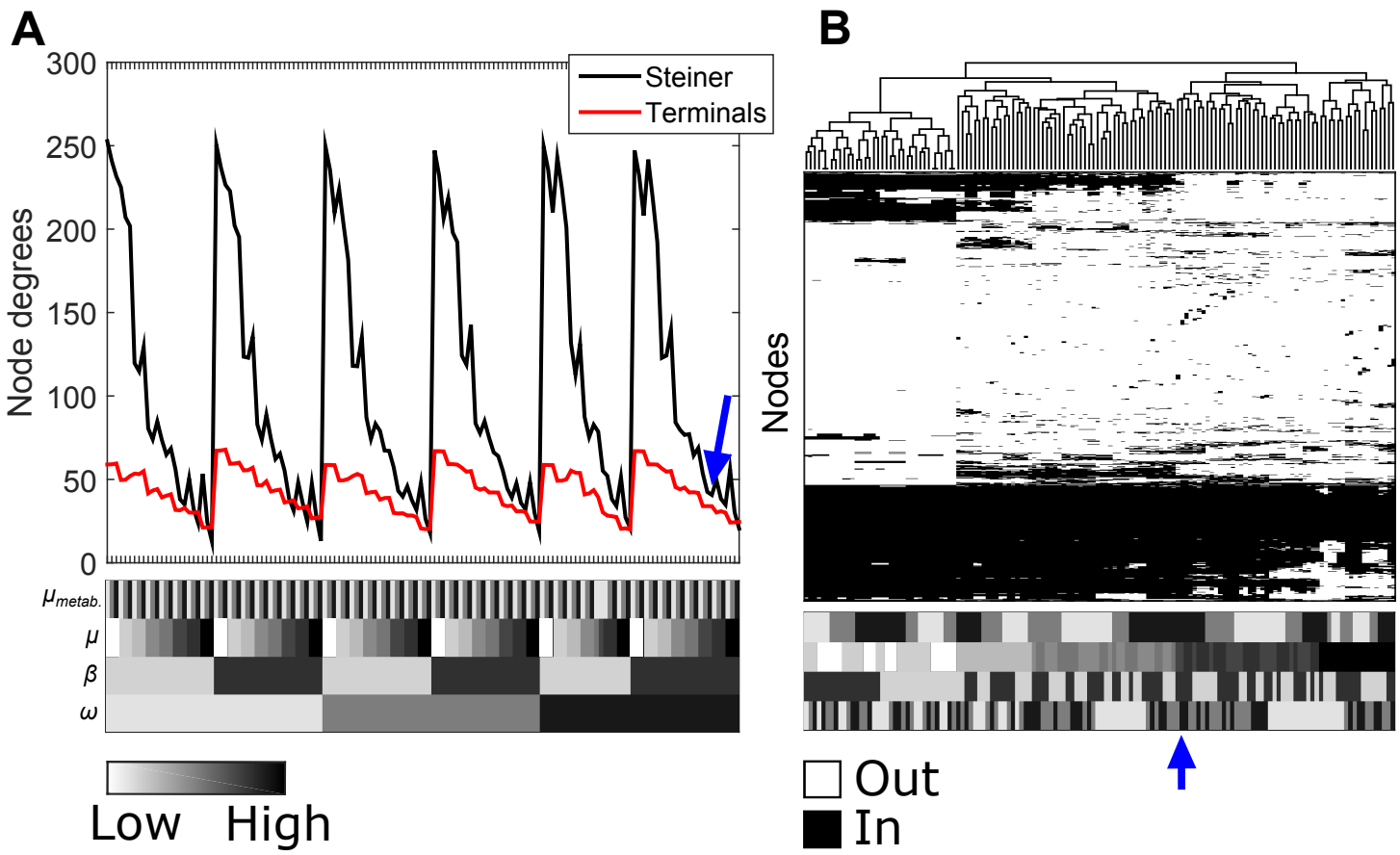


Figure S5. Schematic overview of PCSF algorithm(related to Table 1 and Figure 4).

The PCSF algorithm is initialized by connecting all terminal species (data in box at top right) via an artificial root node. A message-passing algorithm is then run to generate a network model that minimizes the overall objective function (middle box on right, first equation) which balances penalties accrued by excluding data (prize function is in middle box on right, second equation) versus costs required to include edges between nodes (lower confidence edges are more costly). The schematic on the left is a toy representation of a final forest output form one run of the algorithm. The shading in the area with edges from transcription factors to genes indicates that we do not directly include transcription factor-gene edges or gene nodes in the model, though the prize values on transcription factors are influenced by inferred regulation near differentially expressed genes.

Figure S6


A
B
Steiner
Terminals
Node degrees
300
250
200
150
100
50
0
$\mu_{metab.}$
μ
β
ω
Low
High
Nodes
Out
In

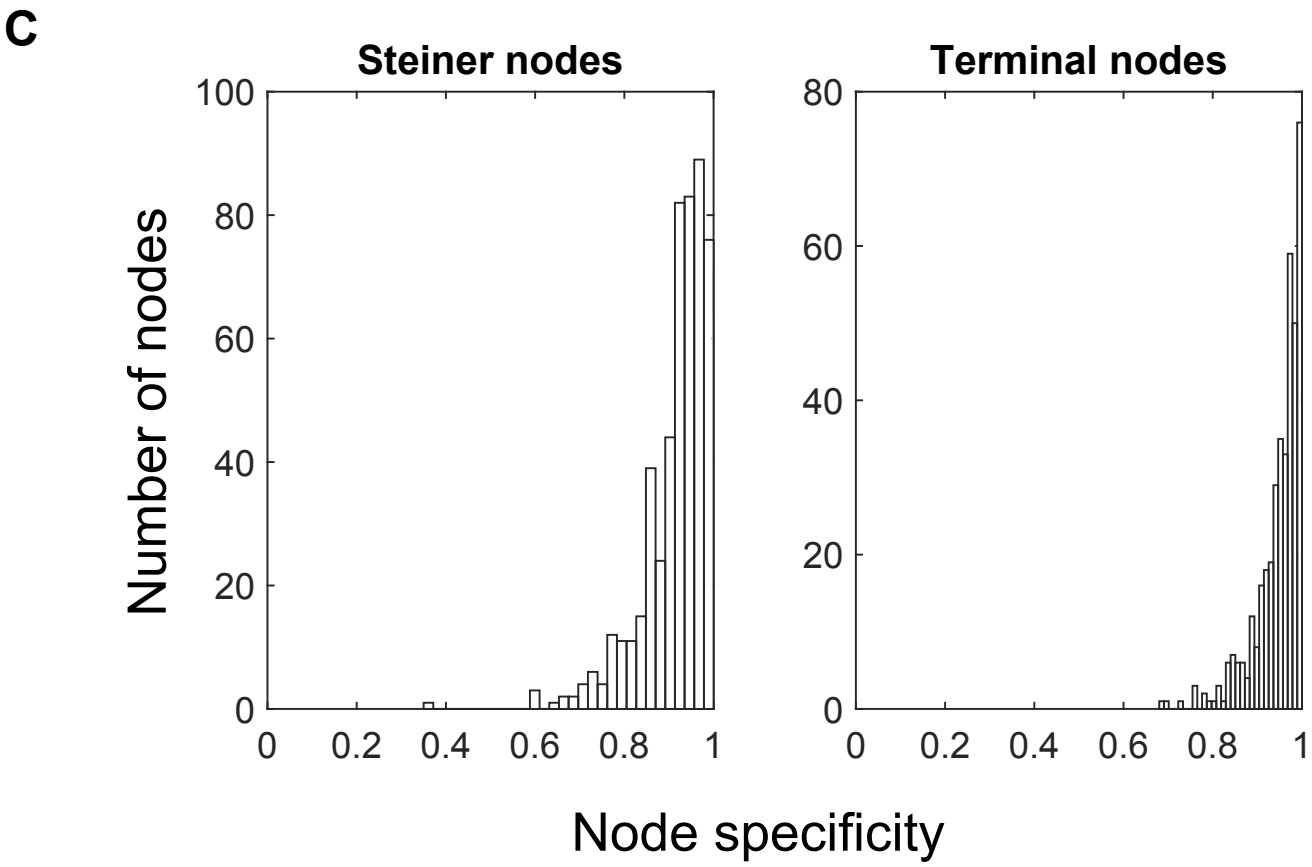
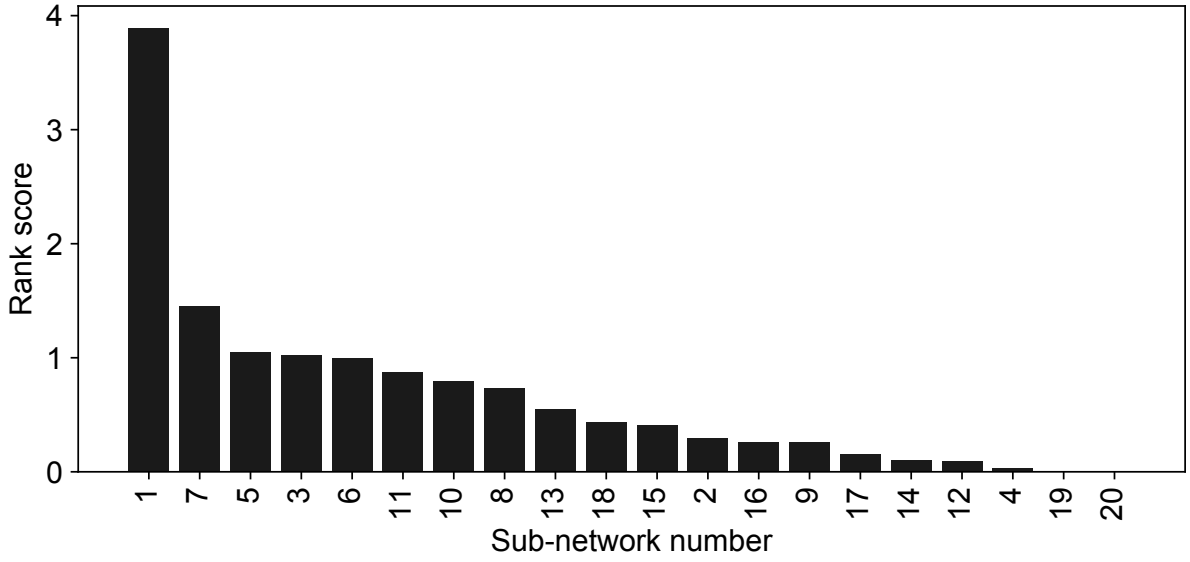

C
Steiner nodes
Terminal nodes
Number of nodes
100
80
60
40
20
0
0
0.2
0.4
0.6
0.8
1
Node specificity

Figure S6. PCSF model parameter selection criteria and final node specificities (related to Figure 4).
We ran the base PCSF algorithm across a number of settings for the major tuning parameters. The plot in (A) shows the average degree of Steiner and terminal nodes included in the optimal solutions at the given parameter settings and (B) shows a binary clustergram of nodes included/excluded in the same solutions. We show tested parameter ranges for: β on [5, 10], μ for metabolite nodes on [1e-5, 1e-4, 1e-3], μ for protein nodes on [1e-7, 5e-7, 1e-6, 5e-6, 1e-5, 5e-5, 1e-4, 5e-4], and ω on [1, 2, 3]. The blue arrows indicate the final parameter settings: β = 10, μ for metabolites = 1e-3, μ for protein nodes = 5e-5, and ω = 3. The final parameter values were chosen based on several criteria: 1) there was a small difference in Steiner node and terminal node degrees (A), 2) very general, "hubby" nodes (e.g. ubiquitin, water, etc.) were excluded from the solution (generally the left-most cluster, top row, in (B), and 3) a high percentage of terminal nodes were included in the solution (note the elimination of several terminal nodes at higher values of μ and lower values of β towards the bottom right corner in (B)). Also note that the most stable clustering of solutions is at the left-most side of the panel (B) clustergram; however, these solutions are generally those that contain the most "hub" nodes and are the most discrepant in terms of node degree differences and were therefore discarded.
(C) We ran the PCSF algorithm at the chosen optimal parameter settings 100 times with random, degree-matched terminals and computed specificities as one minus the fraction of times a node in the optimal solution appeared in the random solutions. Overall, both Steiner and terminal nodes included in our final model are generally highly specific to our particular system.



1. Amino acid/pyruvate metabolism
2. Glucose/glycogen metabolism
3. Endocytosis
4. Unfolded protein response
5. Apoptosis
6. Histidine/tyrosine/retinol metab.
7. Fatty acid/lipid oxidation
8. Arginine/proline metab., TCA cycle
9. Cell-cell interactions
10. ECM
11. Transcription
12. Xenobiotic/drug metabolism
13. Bile acid metabolism
14. Steroid biosynthesis
15. Mitosis
16. Steroid metabolism
17. Cholesterol/lipid metabolism
18. Folate, purine/pyrimidine metab.
19. Monoxygenase activity
20. N/A

Figure S7. PCSF model sub-network rankings(related to Figure 5).

We ranked the PCSF sub-networks by prize density according to: sum(prizes) * (fraction of total network size).

DESCRIPTIONS OF SELECT PCSF MODEL SUB-NETWORKS

Related to Figure 2, Supplementary file 1, and Supplementary data 5

Sub-network 1: Sub-network 1 is the largest (124 nodes) and is most strongly enriched for pathways and processes related to amino acid and amino sugar metabolism and transport. This sub-network is also enriched for species involved in pyruvate metabolism, including the terminals PEP carboxykinase (PCK1) and the pyruvate kinase liver/RBC isozyme (PKLR). The amino acid metabolites in this sub-network are included as both terminals (alanine, asparagine, aspartic acid, cysteine, glycine, and serine) and Steiner nodes (glutamine, lysine, and tyrosine); all of the terminal amino acids, with the exception of alanine, are down-regulated in HFD livers. Protein terminals involved in amino acid metabolism include aminoadipate aminotransferase (AADAT), aminoadipate-semialdehyde synthase (AASS), alanine-glyoxylate aminotransferase 1 and 2 (AGXT, AGXT2), glutaminase 2 (GLS2), glutamic-oxaloacetic transaminase 1 (GOT1), and serine hydroxymethyltransferase 1 (SHMT1). Most of these protein terminals are down-regulated in HFD livers, with the exception of AGXT2 and SHMT1 which increase. Steiner nodes involved in amino acid metabolism in this sub-network include aldehyde dehydrogenase 7A1 (ALDH7A1), asparagine synthetase (ASNS), and glutamine. Several of our high-ranking Steiner nodes are also present in this sub-network, including glyoxylic acid, Glyoxylate/hydroxypyruvate reductase (GRHPR), and CNDP2, or peptidase A.

Sub-network 2: Sub-network 2 is enriched in glucose and glycogen metabolic processes and includes the up-regulated carbohydrate metabolites D-glucose, D-fructose, and D-maltose from our metabolomic data as well as enzymatic regulators of carbohydrate metabolism, including glucokinase (GCK), glucokinase regulator (GCKR), and glycogen debrancher enzyme (AGL), that are up-regulated by HFD in our proteomic data. Notably, hepatic GCK mutations (either activating or inactivating) cause glycemic dysregulation in humans (Agius, 2009) and alteration of its activity is often seen in patients with diabetes (Caro et al., 1995). Pharmacological activators of GCK can lower plasma glucose levels and likely confer their effects via disruption of the GCK-GCKR regulatory interaction captured in our model (Futamura et al., 2006). The full PCSF model included 29 proteins and metabolites annotated to gluconeogenesis and/or glycolysis pathways (of 128 possible from the databases queried).

Sub-network 5: This sub-network is enriched in apoptotic processes and is described in further detail in the main text.

Sub-network 9: This sub-network is enriched in proteins involved in cell-cell interactions and is described in further detail in the main text.

Sub-network 10: This sub-network is enriched in extracellular matrix (ECM) organizational and structural proteins and described in further detail in the main text.

Sub-network 11*:* Sub-network 11 (71 nodes) is enriched in transcriptional processes and contains the majority of the transcription factors from our motif regression procedure. Nuclear hormone receptors are strongly represented in this sub-network, including the terminals RXRα, RXRβ, PPARα, PXR (NR1I2), LXRβ (NR1H2), and VDR. A number of the Steiner nodes identified in this sub-network are also known to regulate processes relevant to insulin resistance and type 2 diabetes. The thyroid hormone receptor (THRA), which also ranks highly in our Steiner node ranking scheme, potentiates insulin signaling in *db/db* mice by enhancing tyrosine phosphorylation of IRS-1 and activation of PI3-kinase (Lin and Sun, 2011). The THRA agonists GC-1 and KB2115 reduce hepatic steatosis in *ob/ob* mice, though they only improve glucose tolerance and insulin sensitivity at high doses that also disrupt thermogenic control (Martagon et al., 2015). Additionally, the Steiner nodes NCOA6 and NCOR2 (aka SMRT) play roles in regulating insulin signaling and sensitivity (Kim et al., 2012; Mottis et al., 2013).

Sub-network 13: Sub-network 13 is enriched in bile acid synthesis and is almost completely composed of protein-metabolite interactions. The PCSF algorithm introduced many Steiner metabolites to connect differential bile acid pathway members from our omic data. Previous studies have identified important roles for bile acid signaling mediated by the nuclear hormone receptor FXR and the guanine nucleotide coupled receptor TGR5 in metabolic syndrome (Chavez-Talavera et al., 2017). Among the terminal nodes in this sub-network are ATP binding cassette B11 (ABCB11), cytochrome P450 proteins 27A1 (CYP27A1) and 7B1 (CYP7B1), very long-chain acyl-CoA synthetase (SLC27A2), and organic anion transporter 1B3 (SLCO1B3).The terminal CYP27A1 is connected to the Steiner metabolite 27-hydroxycholesterol, the product of CYP27A1's enzymatic action on cholesterol in the first step of the alternate bile acid metabolism pathway. CYP7B1 further metabolizes 27-hydroxycholesterol to 7α-hydroxysterol intermediates in this pathway. SLC27A2, another terminal, metabolizes the precursor of cholic acid 3α,7α,12α-trihydroxy-5β-cholestanoic acid (THCA, a Steiner node) to its CoA derivative (THCA-CoA, another Steiner node) in steps leading to formation of taurine- and glycine-conjugated bile acids (Mihalik et al., 2002). Bile acid coenzyme A (BAAT), a high-ranking Steiner node, conjugates these bile acids for biliary excretion (Johnson et al., 1991), and is indeed connected to, among other metabolites, the Steiner nodes taurocholic acid and glycocholic acid. The terminal ABCB11 exports bile salts from hepatocytes (Garzel et al., 2014) and SLCO1B3, a liver-specific organic anion influx transporter, transports bile salts, thyroid hormones, and eicosanoids (Ohtsuka et al., 2006). Both of these terminals are connected by Steiner bile acids and bile acid conjugates in the PCSF network. Also included as a Steiner protein is ACSL1, an acyl-CoA synthetase that plays a role in lipid biosynthesis and fatty acid degradation, which ranks highly in our scheme and connects several fatty acid metabolite terminals.

SUPPLEMENTARY EXPERIMENTAL PROCEDURES

mRNA-Seq analysis

We prepared mRNA-Seq libraries from three CD and three 16 week HFD mouse livers using the TruSeq RNA Sample Prep Kit v1 (Illumina) and size-selected using 2% agarose gel electrophoresis for 180 +/- 25 base-pairs of insert. We multiplexed mRNA-Seq libraries and paired-end sequenced samples for 40-50 base-pairs on an Illumina Hi-Seq 2000 machine. On average, we obtained ~20-30 million raw paired-end sequencing reads. The reads were aligned to known mouse RefSeq gene transcripts obtained from the UCSC table browser (Karolchik et al., 2004) (accessed on January 25, 2012) and the mouse genome (build mm9) with the splice junction-aware short-read alignment tool TopHat (version 1.4.0) (Trapnell et al., 2009). We restricted TopHat to only align to known transcript splice junctions. We observed strong intra-sample correlations between CD (Pearson's $r > 0.995$) and HFD ($r > 0.993$) replicate gene read count levels (**Figure S2A**). We used the Bioconductor package conditional quantile normalization (CQN, version 1.6.0) (Hansen et al., 2012) to remove systematic biases due to GC-content and gene length coverage and used DESeq2 (version 1.0.18) (Love et al., 2014) to perform differential expression analyses. We considered a gene to be differentially expressed if it possessed an absolute $\log_2$ fold-change between conditions ≥ 0.5, an FDR-adjusted p-value (q-value) ≤ 0.05, and was expressed in at least one tested condition (i.e. ≥ 0.1 FPKM).

Differential ChIP-Seq peak analysis

We performed differential peak analyses independently for each histone mark between CD and HFD samples. We used MACS-called peaks (version 1.4.0rc2) from replicate-pooled runs possessing a p-value < 1e-6 and an FDR < 10% and merged significant peak regions from each condition into one set of common genomic loci using BEDTools (Quinlan and Hall, 2010). We then extracted raw read counts in these regions from each individual replicate ChIP-Seq sample and used DESeq2 (Love et al., 2014) to perform the differential enrichment analyses on the read counts. We considered regions possessing an FDR-corrected p-value < 0.05 as significant.

Global proteomics data acquisition and analysis details

We collected global proteomic data from four CD and four 16 week HFD mouse livers. Liver powder was homogenized (Polytron) in ice-cold lysis buffer consisting of 8M urea supplemented with 1 mM sodium orthovanadate, 0.1% Nonident P-40 (NP-40), and protease inhibitor and phosSTOP tablets (Roche). Samples were homogenized on ice using 5x10 sec pulses, with 10 sec intervening periods to prevent tissue heating. Protein concentrations were quantified by a bicinchoninic acid (BCA) assay (Pierce). Homogenized liver samples were reduced in 10 mM DTT at 56°C for 45 min and alkylated with 50 mM iodoacetamide at room temperature for 1 hour in the dark. Proteins were digested to peptides with sequencing grade trypsin (Promega) at 1:100 enzyme to

substrate ratio at room temperature overnight in 100 mM ammonium acetate, pH 8.9. Trypsin activity was then quenched with acetic acid at a final concentration of 10%. Urea was removed by reverse-phase desalting using C18 cartridges (Waters). Samples were then lyophilized and stored at -80°C. Peptides were labeled with iTRAQ 8plex isobaric mass tags (iTRAQ, AB Sciex) according to the manufacturer's protocol.

The iTRAQ labeled peptides were analyzed by multidimensional LC-MS/MS (DEEP SEQ mass spectrometry) as described previously (Zhou et al., 2013). Briefly, a NanoAcquity UPLC system (Waters, Milford, MA) utilized 2 binary pumps, an autosampler, and an additional 6-port, 2-position valve (Valco, Austin, TX). Peptides were first fractionated at high pH (10.0) using a reversed phase column (200 µm ID fused silica x 20 cm packed with 5 µM XBridge C18). In the second dimension, peptides were further resolved at high pH by strong anion exchange chromatography (200 µm fused silica x 20 cm packed with 5 µM SAX; SEPAX technologies, Newark, DE). Peptides were eluted from each dimension using solutions of acetonitrile and/or ammonium formate (pH 10) for a total of 20 fractions. Peptides from each fraction were trapped on the final dimension precolumn (200 µm ID fused silica x 4 cm of POROS 10R2) after in-line dilution with 0.1% formic acid, and subsequently resolved on an analytical column (25 µm ID fused silica packed with 100 cm of 5 µm Monitor C18, Column Engineering, Ontario, CA) using an organic gradient: 2–50% B in 580 min, A=0.1% formic acid, B=acetonitrile with 0.1% formic acid. The analytical column terminated with a ~1 µm diameter electrospray emitter (Ficarro et al., 2009), positioned near the mass spectrometer orifice (5600 Triple TOF mass spectrometer, ABI, Framingham, MA) by use of a computer controlled Digital Picoview Platform (New Objective, Woburn, MA). The 5600 Triple TOF was operated in information dependent mode (IDA), with the top 50 precursors (charge state +2 to +5, >70 counts) in each MS scan (800 ms, scan range 350-1500 m/z) subjected to MS/MS (minimum time 140 ms, scan range 100-1400m/z). Dynamic exclusion was enabled, exclusion duration 20 seconds, and the isolation window was set to unit resolution. Electrospray voltage was set to 2.2 kV.

Raw mass spectrometry data files were searched using Protein Pilot V4.4 (AB Sciex, Framingham, MA), with parameters specifying trypsin digestion, 8-plex labeling of peptides, and carbamidomethylation of cysteine residues. Multiplierz scripts (Parikh et al., 2009) were used to filter PSMs to a 1% false discovery rate, extract iTRAQ reporter ion intensities, and correct for isotopic impurities as well as minor variations in source protein concentration. We re-aligned peptides to the set of non-redundant protein sequences (filtered for mouse sequences) obtained from NCBI's BLAST FTP database (accessed August 5, 2013) using the BLAST command line tool with the recommended parameters for short amino acid sequence alignment: -p blastp -e 200000 -F F -G 9 -E 1 -M PAM30 -W 2 -A 40 -C F.

We used a weighted least squares (WLS) regression procedure to find differentially expressed global proteins between conditions (based on (Oberg and Mahoney, 2012)). We included only peptides that uniquely mapped to one protein and ignored isoform-specific information. We first performed a sample-wise normalization on the isotope-corrected iTRAQ channels using the procedure described in (Anders and Huber, 2010) to adjust for global

differences in overall protein abundance and imputed missing values with a k-nearest-neighbors procedure (k = 10). We observed strong intra-sample correlations between CD (Pearson's r > 0.946) and HFD (r > 0.921) replicates using these normalized and imputed values (**Figure S2B**). We then assigned weights to individual peptide measurements per condition by fitting a locally weighted curve through a plot of the coefficient of variation (CV) versus $\log_2$ mean abundance for all peptides. The CV for each peptide in each condition was set to the maximum of the fit value or the raw calculated CV value and weights were assigned as the inverse of this value. For every protein *i*, we fit a WLS regression model that included information from all peptides 1…*j* to estimate the overall effect of the treatment condition on the expression level of the protein, i.e.:

$$\log_2(y_{ijcr}) = \mu_i + pep_{ij} + cond_{ic} + \varepsilon_{ijcr}$$

where y_{ijcr} are the corrected abundances (iTRAQ intensities) from replicate measurements *r* of peptides *j* derived from protein *i* in conditions *c* (i.e. CD and HFD), μ_i is the overall fit mean for protein *i*, the pep_{ij} terms are the fit mean abundances for peptides *j* (necessary for aligning distributions as individual peptide abundances from the same protein can vary over orders of magnitude), $cond_{ic}$ is the overall treatment effect on protein *i*, and ε_{ijcr} are the error terms. The $cond_{ic}$ terms are of interest as these are the fit $\log_2$ fold-change values for the proteins between the conditions. We tested the $cond_{ic}$ terms for significance using two-tailed t-tests and corrected p-values for multiple hypothesis testing using the Benjamini-Hochberg false discovery rate (FDR) procedure for all *i* proteins tested. The *lscov* function in MATLAB (The MathWorks, Inc., Natick, MA) was used to implement the WLS regression procedure. We retained as significantly differentially expressed proteins those with an FDR-corrected p-value < 0.1. The data used for the weighted least squares regression procedure and results are provided in **Supplementary data 1**.

Motif regression details

Histone valleys: We scanned enriched H3K4me1, H3K4me3, and H3K27Ac regions for histone valleys, or areas of local signal depletion in broad enrichment regions. We used peaks called from pooled-replicate runs of MACS (Zhang et al., 2008) against IgG controls for each ChIP-Seq data type. We created smoothed signal profiles for these peaks from the aligned sequencing reads for each dataset. To do this, we shifted reads a fixed distance towards their 3' ends (by the amount estimated from the MACS peak-shift model), created profiles from the read pileup data in the peak regions, and smoothed these profiles using a moving average filter. We then used a numerical procedure to find local minima in each signal profile. For each base-pair in an individual enrichment region, a "valley score" was calculated as the difference in read pileup height between the minimum of the two neighboring local maxima in the +/- 500 base-pair windows around the current point and the current point pileup height itself. A point whose valley score was 50% smaller in magnitude than the smaller of the two nearest local maxima was designated as a valley point. Neighboring valley points were then merged into a single region and the point with the maximum valley score in this region was reported as the valley location. We then took a fixed 100 base-pair window around each valley

location and reported these as the valley regions for each dataset. We then created a combined set of valleys from all the discovered regions in the H3K4me1, H3K4me3, and H3K27Ac peak regions, considering one base-pair overlap as a valid intersection while also retaining unique regions from each individual dataset. In total, we found 123,974 unique valley regions that were used for further analyses.

Motif matching and scoring in valleys: We used 1,588 DNA-binding motifs annotated to human and mouse transcriptional regulatory proteins from release 2013.3 of TRANSFAC® (Wingender et al., 1996), represented as position-specific scoring matrices (PSSMs). We extracted the underlying genomic sequences from the histone valley regions and used TAMO (Gordon et al., 2005) to store the motif PSSMs, read in the valley region sequences, and score the sequences for matches to the motifs. We computed a normalized log-likelihood ratio (LLR) score as $LLR_{norm} = (LLR - LLR_{min})/(LLR_{max} - LLR_{min})$ for every *k*-base-pair sub-sequence in the valley regions, where *k* is the length of the motif PSSM. A motif match was called if LLR_{norm} was greater than or equal to the TRANSFAC®-computed minimum false positive matrix similarity score threshold (minFP) for that motif. The maximum matching LLR_{norm} for each motif in each sequence was retained and used to create a matrix of genomic regions by motifs. Regions with no matches to a given motif were given a score of zero.

Transcription factor affinity score calculations: We retained histone valley regions that were within -50/+10 kilobases from the transcription start sites of at least one differentially expressed gene between CD and HFD livers. We computed transcription factor affinity (TFA) scores for each motif against each gene as:

$$TFA_{m,g} = \sum_{i=1}^{n} LLR_{m,g,i} \cdot e^{-d_{m,g,i}/d_o(m)}$$

where $TFA_{m,g}$ is the TFA score for motif *m* against gene *g*, $LLR_{m,g,i}$ is the normalized LLR score in the i^{th} valley region near gene *g* for motif *m*, $d_{m,g,i}$ is the distance of the i^{th} motif match from the TSS of gene *g* in base-pairs, and $d_o(m)$ is the exponential distance constant for motif *m* (set to 10,000 bases for all motifs here). Transfer functions incorporating the number of binding sites and distances of these from genes perform well when predicting regulator influences on gene expression (Sikora-Wohlfeld et al., 2013). From here, a matrix of genes by TFA scores was created for the set of differentially expressed genes between CD and HFD.

Motif regression: For each transcription factor motif, we built a simple univariate linear regression model to predict gene expression levels from the computed TFA scores. We used the individual TFA scores for each gene as the predictor variables and the corresponding $\log_2$ FPKM expression values as the response variables. We mean-centered and variance scaled (i.e. z-scored) both the predictor and response data and assessed the significance level for each regression slope using the t-distribution. We corrected individual p-values for multiple hypotheses and retained significant motifs as those possessing an FDR q-value < 0.01. By this metric, 358 motifs possessed significant regression slopes. The data and code used to perform this procedure are provided in **Supplementary data 3**.

PCSF modeling details

PCSF formulation: The prize-collecting Steiner forest (Tuncbag et al., 2013; Tuncbag et al., 2016) aims to find a forest $F(V_F, E_F)$ from the graph $G(V, E, c(e), p(v))$, with nodes V, edges E, edge costs $c(e) \geq 0$, and node prizes $p(v)$ for $v \in V$, that minimizes the objective function:

$$PCSF(F) = \sum_{v \notin V_F} p(v) + \sum_{e \in E_F} c(e) + \omega \cdot \kappa$$

where κ is the number of trees in the forest, ω is a tuning parameter that influences the number of trees included in the final forest, and:

$$p(v) = \beta \cdot p_o(v) - \mu \cdot \text{degree}(v)^n .$$

The β parameter scales the importance of node prizes versus edge costs in the optimization and can be a uniform value for all terminals or uniquely set for a given input data type. We employed a "negative prize" scaling scheme to each node in G proportional to its degree, or number of connections in the graph, to reduce the influence of highly-connected, well-studied nodes that have a high likelihood of appearing in PCSF solutions run with almost any input data. The parameter μ scales the influence of the negative prizes and the exponent n allows for non-linearity in the scaling.

Interactome: We built a combined protein-protein and protein-metabolite interactome from which the PCSF derived connections between our input data. We used the set of human interactions contained in version 13 of the iRefIndex database (Razick et al., 2008) as our source for protein-protein interactions, which consolidates information from a variety of source databases. We used the MIscore system (Villaveces et al., 2015) to assign confidence scores (ranging from 0 to 1) to these interactions, which considers the number of publications (publication score), the type of interaction (type score), and the experimental method used to find the interaction (method score). We extracted the relevant scoring information for interactions from the iRefIndex MITAB2.6 file, using the redundant interaction group identifier (RIGID) to consolidate interactions between the same two proteins reported by multiple databases, and used a java implementation of MIscore (version 1.3.2, obtained from https://github.com/EBI-IntAct/miscore/blob/wiki/api.md) with default parameters for individual score weights. We only considered interactions between two human proteins (i.e. we excluded human-viral interactions) and converted protein identifiers, generally provided as UniProt or RefSeq accessions, to valid HUGO gene nomenclature committee (HGNC) symbols. Once converted, we removed redundant interactions (generally arising from isoform-specific

interactions that map to the same protein/gene symbols) and retained the maximum observed score. This produced a total of 175,854 unique protein-protein interactions.

We collected protein-metabolite interactions from version 3.6 of the human metabolome database (HMDB) (Wishart et al., 2013) and supplemented these with additional manually curated interactions from the human metabolic reconstruction Recon 2 (version 3) (Thiele et al., 2013). We assigned uniform weights to the HMDB interactions, using the median of protein-protein interaction scores (~0.448) as this value, and protein identifiers were converted to valid HGNC gene symbols. We extracted reaction-gene link information from the Recon 2 MATLAB file Recon2.v03.mat, available at humanmetabolism.org. For edges included in both HMDB and Recon 2, we added 0.2 or 0.3 to their default edge scores if they were assigned a curated score of 3 or 4 in Recon 2. We included unique edges from Recon 2 with a curated score of 2 or greater, adding 0.1 or 0.3 to their edge scores if they were assigned a score of 3 or 4. We also excluded edges between drugs, drugs metabolites, and metabolites of non-endogenous origins (according to HMDB's origin information). This protein-metabolite interactome was merged with the iRefIndex protein-protein interactions to produce a final interaction network of 1,016,322 edges between 36,891 nodes.

PCSF run details and final model selection: We converted all mouse genes (proteins) to their human orthologs using orthology information from the mouse genome informatics (MGI) database and HGNC for proteins. Also, we retained metabolites that mapped to a valid HMDB identifier. We used as prize values in the PCSF optimization the absolute values of the $\log_2$ fold-changes between CD and HFD livers for the global protein and metabolites terminals and used the absolute values of the regression coefficients from the motif regression results for the transcription factor terminals. In total, we supplied to the PCSF 83 metabolites that possess valid HMDB identifiers, 329 global proteins that were successfully mapped to orthologous human genes, and the 272 transcriptional regulators identified by our motif regression analysis.

We ran the PCSF across an array of values for the relevant tuning parameters. The β values for global proteins and metabolites varied over [5, 10] and for transcription factors over [1, 5, 10]. The ω parameter was varied over [1, 2, 3] and the μ for protein terminals (global proteins and transcription factors) varied over [1e-7, 5e-7, 1e-6, 5e-6, 1e-5, 5e-5, 1e-4, 5e-4]. The μ for metabolite terminals was varied over [1e-5, 1e-4, 1e-3]. The exponent n used in the negative prize scaling was set to 2 for all PCSF runs. We elected to use different β values for the transcription factor termini as the prize values for these are on a different scale from the global proteins and metabolites (regression coefficients versus observed $\log_2$ fold-changes). We also used independent μ values for metabolites and proteins as the degree distributions between the two source databases differ. The message-passing algorithm used to solve the PCSF problem requires as additional input a depth parameter D, which specifies the maximum path length from the artificial source node to any node in the forest, a reinforcement parameter g, which influences the convergence of the solution by producing more optimal solutions at lower values at the expense of increased run time, and a noise parameter r, which adds random noise to edges during run time. We used D values of 5, 7, and 10, a g of 1e-3 (to

force more optimal solutions), and an r of 1e-5. For sub-optimal runs (see below), this r value was manipulated to add noise to the interactome edge scores.

We used several selection criteria to arrive at a final set of parameters for our network model. Generally, we preferred larger values for β in an effort to include as much data as possible in the final networks. The μ parameters strongly influences the type of nodes included in the solution. To select appropriate μ values, we visualized, for every combination of all the tuning parameters, both the average degrees of selected terminal and Steiner nodes, as well as the nodes included in the optimal solutions (see **Figure S6A-B**). We preferred solutions in which the difference in average degree distribution between the terminal and algorithm-introduced Steiner nodes were similar; a large disparity here is indicative of generic solutions in which very common, high degree nodes are used to connect the data. We also clustered a binary heatmap of nodes included in the various solutions to look for parameter regimes in which the included nodes at least partially stabilize. The most stable region occurs at low μ values for both proteins and metabolites; however, it is in these regions where we see the most generic Steiner nodes included (e.g. ubiquitin and amyloid beta precursor for proteins and water, oxygen, and NADH for metabolites) and where the degree distributions are very discrepant between terminal and Steiner nodes. Therefore, we selected a solution in a semi-stable region of included nodes that excluded most of these generic species and where the Steiner and terminal node degree separation was not too large. We also employed robustness and specificity tests of results to our particular input data to validate our parameter selections.

The final PCSF solution presented here used β = 10 for global proteins and metabolites, β = 1 for transcription factors, ω = 3, D = 7, μ = 5e-5 for global proteins and transcription factors, and μ = 1e-3 for metabolites. We merged the optimal PCSF run with 50 "sub-optimal" solutions run at the same parameter settings but with random noise added to the interactome edge weights. Edge noise was introduced via the r parameter in the message-passing code and was set to 0.1. This particular randomization procedure allows the algorithm to find alternative connections between the data nodes that may be of biological importance when interpreting the data and network results. Note that the PCSF model is no longer a forest due to the merger of random runs. We also re-introduced all available edges between included solution nodes from the interactome, whether selected by the algorithm or not, into the final solution.

We assessed the specificity of each node in the final solution by running the algorithm 100 times at the same parameter settings, but with random input data. We define specificity in this context as the frequency with which a given terminal or Steiner node in the final PCSF model appears in runs with random input data. For each random run, we selected random terminals matching the degree distribution of the real terminals. For each real terminal, a node from the interactome matching (within a small error range) the degree of the original true terminal was randomly selected and assigned the same prize value as the original true terminal. At these final parameter settings, we found that the included terminal and Steiner nodes were generally highly specific to our particular data (see **Figure S6C**).

The full PCSF solution includes 76%, 91%, and 13% of includable (i.e. species converted to appropriate identifiers or genes/proteins possessing a human ortholog) metabolites, global proteins, and transcription factors. The low inclusion percentage of transcriptional regulators results from the fact that the β value for these terminals is lower than the β value applied to the metabolites and global proteins. When solutions were run with larger β values, a greater percentage of the transcriptional regulators were indeed included in the models (74% at $\beta_{TF} = 5$ and 86% at $\beta_{TF} = 10$). The solutions run with these larger values generally produced networks with large sub-clusters comprised almost completely of interconnections amongst the transcriptional regulators themselves. Therefore, we selected a smaller value to highlight the most influential transcriptional regulators.

PCSF model clustering and visualization. We used a community clustering algorithm that maximizes network modularity (Blondel et al., 2008) to break the full PCSF model into smaller sub-networks. We performed enrichment analyses on the nodes in each of these subnetworks using as gene sets human gene ontology terms, pathway members from the small molecule pathway database (SMPDB) (Jewison et al., 2014), and canonical pathway node sets provided by MSigDB (Subramanian et al., 2005) derived from various database sources. We visualized all networks with Cytoscape (Shannon et al., 2003).

PCSF model node ranking scheme. We used a weighted sum of feature scores to rank nodes contained in our PCSF network. We did this for Steiner and terminal nodes separately as some feature score distributions differ between the two node sets. The features we used for each node were: 1) the robustness of the node to edge noise (i.e. the frequency of solutions run with random edge noise including this node), 2) the specificity of the node (one minus the frequency with which the node showed up in solutions run with random input data), 3) the mean of nearest neighbor node specificities ($k = 2$ for all neighborhood features), 4) the mean of interactome edge weights connecting the nearest neighbor nodes, 5) the fraction of neighbor nodes that are terminals, and 6) and the size of the local neighborhood to which the node belongs, which was scored as a saturating function of the neighborhood size. We set the weights for the six feature scores by taking the inverse of the score variances across all nodes; this scheme favors features that have more discriminatory power between nodes. For example, the robustness feature had a higher weight than specificity as the majority of included nodes were highly specific to our problem. The final weights were adjusted such that the sum of the weights equaled one.

TUNEL imaging analysis with CellProfiler

We used CellProfiler (version 2.1.1) (Carpenter et al., 2006) with a custom-built analysis pipeline from modules included in the program to analyze TUNEL images. All images across CD and HFD samples were analyzed in a single run of the program at the same settings. The pipeline we used: 1) loaded images (two channel images for all fields of view, red for DAPI and green for TUNEL), 2) converted images to grayscale, 3) identified nuclei by DAPI staining using the "IdentifyPrimaryObjects" module, setting the typical diameter minimum and maximum to 5 and

25, discarding objects touching the border, using the automatic thresholding strategy, and using shape to distinguish clumped objects and to draw dividing lines between objects, 4) identified TUNEL positive objects with "IdentifyPrimaryObjects" with same settings for DAPI, though we set the minimum and maximum diameter to 4 and 20, 5) used the "RelateObjects" module with nuclei treated as parents and TUNEL objects treated as children, and 6) used "FilterObjects" to filter nuclei by TUNEL positive objects. The TUNEL positive percentage per field of view was calculated as the number of positive nuclei over the total. For each liver, we calculated a single TUNEL positive fraction by dividing the total number of TUNEL positive nuclei by the total number of nuclei across all fields of view (n = 9, 7, and 4 for HFD livers; n = 4, 5, and 5 for CD livers). We used a two-tailed t-test to test for statistical significance between CD and HFD livers.

SUPPLEMENTARY REFERENCES

Agius, L. (2009). Targeting hepatic glucokinase in type 2 diabetes: weighing the benefits and risks. Diabetes *58*, 18-20.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol *11*, R106.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J Stat Mech *2008*, P10008-P10008.

Caro, J.F., Triester, S., Patel, V.K., Tapscott, E.B., Frazier, N.L., and Dohm, G.L. (1995). Liver glucokinase: decreased activity in patients with type II diabetes. Horm Metab Res *27*, 19-22.

Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., Moffat, J., et al. (2006). CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol *7*, R100.

Chavez-Talavera, O., Tailleux, A., Lefebvre, P., and Staels, B. (2017). Bile Acid Control of Metabolism and Inflammation in Obesity, Type 2 Diabetes, Dyslipidemia, and Nonalcoholic Fatty Liver Disease. Gastroenterology.

Ficarro, S.B., Zhang, Y., Lu, Y., Moghimi, A.R., Askenazi, M., Hyatt, E., Smith, E.D., Boyer, L., Schlaeger, T.M., Luckey, C.J., et al. (2009). Improved electrospray ionization efficiency compensates for diminished chromatographic resolution and enables proteomics analysis of tyrosine signaling in embryonic stem cells. Anal Chem *81*, 3440-3447.

Futamura, M., Hosaka, H., Kadotani, A., Shimazaki, H., Sasaki, K., Ohyama, S., Nishimura, T., Eiki, J.-I., and Nagata, Y. (2006). An allosteric activator of glucokinase impairs the interaction of glucokinase and glucokinase regulatory protein and regulates glucose metabolism. J Biol Chem *281*, 37668-37674.

Garzel, B., Yang, H., Zhang, L., Huang, S.M., Polli, J.E., and Wang, H. (2014). The role of bile salt export pump gene repression in drug-induced cholestatic liver toxicity. Drug Metab Dispos *42*, 318-322.

Gordon, D.B., Nekludova, L., McCallum, S., and Fraenkel, E. (2005). TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs. Bioinformatics *21*, 3164-3165.

Hansen, K.D., Irizarry, R.A., and Wu, Z. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics *13*, 204-216.

Jewison, T., Su, Y., Disfany, F.M., Liang, Y., Knox, C., Maciejewski, A., Poelzer, J., Huynh, J., Zhou, Y., Arndt, D., et al. (2014). SMPDB 2.0: big improvements to the Small Molecule Pathway Database. Nucleic Acids Res *42*, D478-484.

Johnson, M.R., Barnes, S., Kwakye, J.B., and Diasio, R.B. (1991). Purification and characterization of bile acid-CoA:amino acid N-acyltransferase from human liver. J Biol Chem *266*, 10227-10233.

Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. (2004). The UCSC Table Browser data retrieval tool. Nucleic Acids Res *32*, D493-496.

Kim, G.H., Lee, K.J., Oh, G.S., Yoon, J., Kim, H.W., and Kim, S.W. (2012). Regulation of hepatic insulin sensitivity by activating signal co-integrator-2. Biochem J *447*, 437-447.

Lin, Y., and Sun, Z. (2011). Thyroid hormone potentiates insulin signaling and attenuates hyperglycemia and insulin resistance in a mouse model of type 2 diabetes. Br J Pharmacol *162*, 597-610.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 550.

Martagon, A.J., Lin, J.Z., Cimini, S.L., Webb, P., and Phillips, K.J. (2015). The amelioration of hepatic steatosis by thyroid hormone receptor agonists is insufficient to restore insulin sensitivity in ob/ob mice. PLoS One *10*, e0122987.

Mihalik, S.J., Steinberg, S.J., Pei, Z., Park, J., Kim, D.G., Heinzer, A.K., Dacremont, G., Wanders, R.J.A., Cuebas, D.A., Smith, K.D., et al. (2002). Participation of two members of the very long-chain acyl-CoA synthetase family in bile acid synthesis and recycling. J Biol Chem *277*, 24771-24779.

Mottis, A., Mouchiroud, L., and Auwerx, J. (2013). Emerging roles of the corepressors NCoR1 and SMRT in homeostasis. Genes Dev *27*, 819-835.

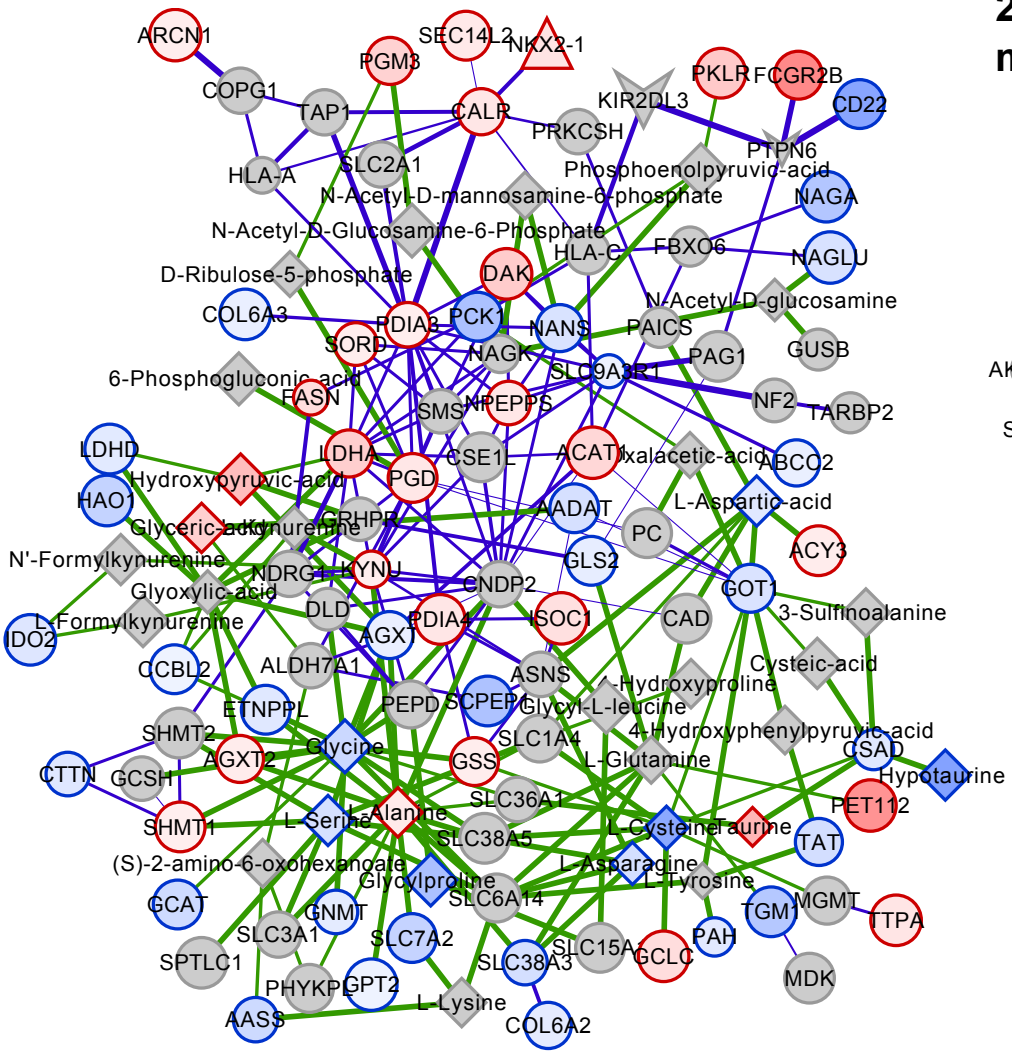
Oberg, A.L., and Mahoney, D.W. (2012). Statistical methods for quantitative mass spectrometry proteomic experiments with labeling. BMC Bioinformatics *13 Suppl 16*, S7.

Ohtsuka, H., Abe, T., Onogawa, T., Kondo, N., Sato, T., Oshio, H., Mizutamari, H., Mikkaichi, T., Oikawa, M., Rikiyama, T., et al. (2006). Farnesoid X receptor, hepatocyte nuclear factors 1alpha and 3beta are essential for transcriptional activation of the liver-specific organic anion transporter-2 gene. J Gastroenterol *41*, 369-377.

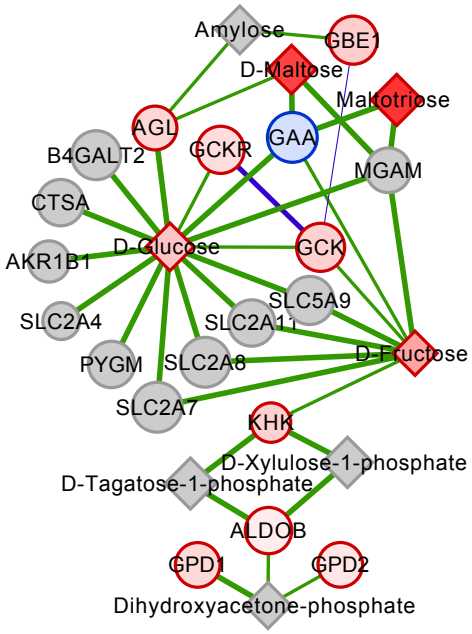
Parikh, J.R., Askenazi, M., Ficarro, S.B., Cashorali, T., Webber, J.T., Blank, N.C., Zhang, Y., and Marto, J.A. (2009). multiplierz: an extensible API based desktop environment for proteomics data analysis. BMC Bioinformatics *10*, 364.
Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.
Razick, S., Magklaras, G., and Donaldson, I.M. (2008). iRefIndex: a consolidated protein interaction database with provenance. BMC Bioinformatics *9*, 405.
Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res *13*, 2498-2504.
Sikora-Wohlfeld, W., Ackermann, M., Christodoulou, E.G., Singaravelu, K., and Beyer, A. (2013). Assessing computational methods for transcription factor target gene identification based on ChIP-seq data. PLoS Comput Biol *9*, e1003342.
Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545-15550.
Thiele, I., Swainston, N., Fleming, R.M., Hoppe, A., Sahoo, S., Aurich, M.K., Haraldsdottir, H., Mo, M.L., Rolfsson, O., Stobbe, M.D., et al. (2013). A community-driven global reconstruction of human metabolism. Nat Biotechnol *31*, 419-425.
Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.
Tuncbag, N., Braunstein, A., Pagnani, A., Huang, S.S., Chayes, J., Borgs, C., Zecchina, R., and Fraenkel, E. (2013). Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem. J Comput Biol *20*, 124-136.
Tuncbag, N., Gosline, S.J., Kedaigle, A., Soltis, A.R., Gitter, A., and Fraenkel, E. (2016). Network-based interpretation of diverse high-throughput datasets through the Omics Integrator software package. PLoS Comput Biol *(In press)*.
Villaveces, J.M., Jimenez, R.C., Porras, P., Del-Toro, N., Duesbury, M., Dumousseau, M., Orchard, S., Choi, H., Ping, P., Zong, N.C., et al. (2015). Merging and scoring molecular interactions utilising existing community standards: tools, use-cases and a case study. Database (Oxford) *2015*.
Wingender, E., Dietze, P., Karas, H., and Knuppel, R. (1996). TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res *24*, 238-241.
Wishart, D.S., Jewison, T., Guo, A.C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., et al. (2013). HMDB 3.0--The Human Metabolome Database in 2013. Nucleic Acids Res *41*, D801-807.
Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol *9*, R137.
Zhou, F., Lu, Y., Ficarro, S.B., Adelmant, G., Jiang, W., Luckey, C.J., and Marto, J.A. (2013). Genome-scale proteome quantification by DEEP SEQ mass spectrometry. Nat Commun *4*, 2171.

Data S6 (page 1)

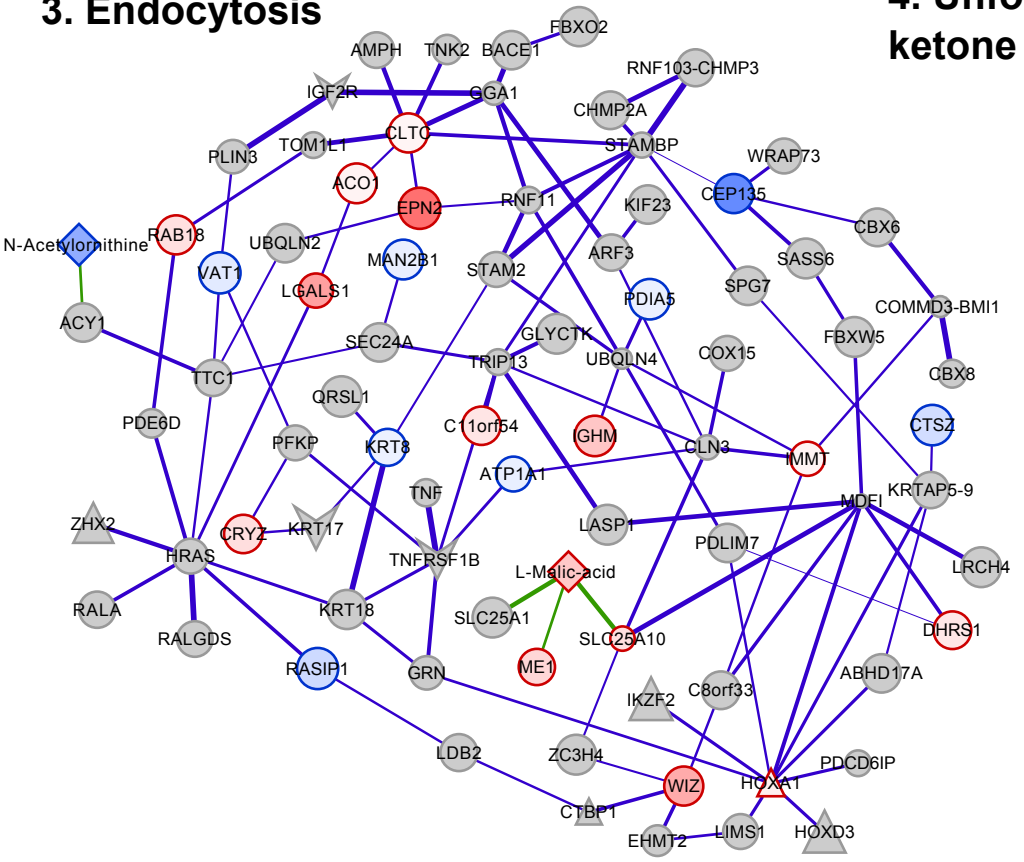
1. Amino acid/pyruvate metabolism



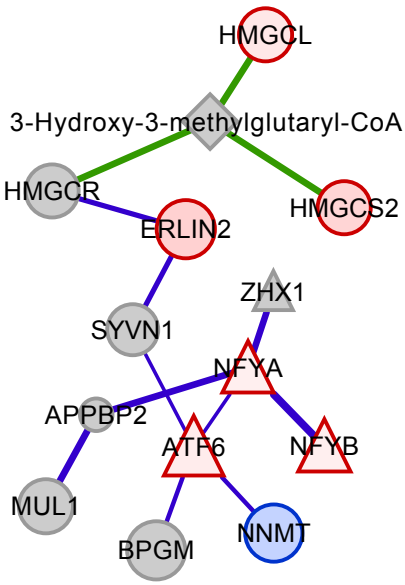
2. Glucose/glycogen metabolism



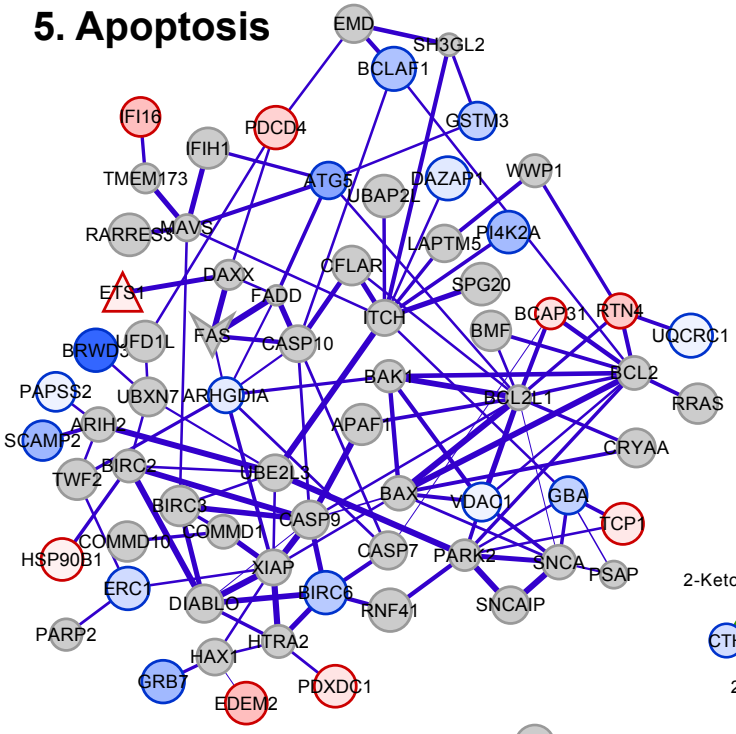
3. Endocytosis



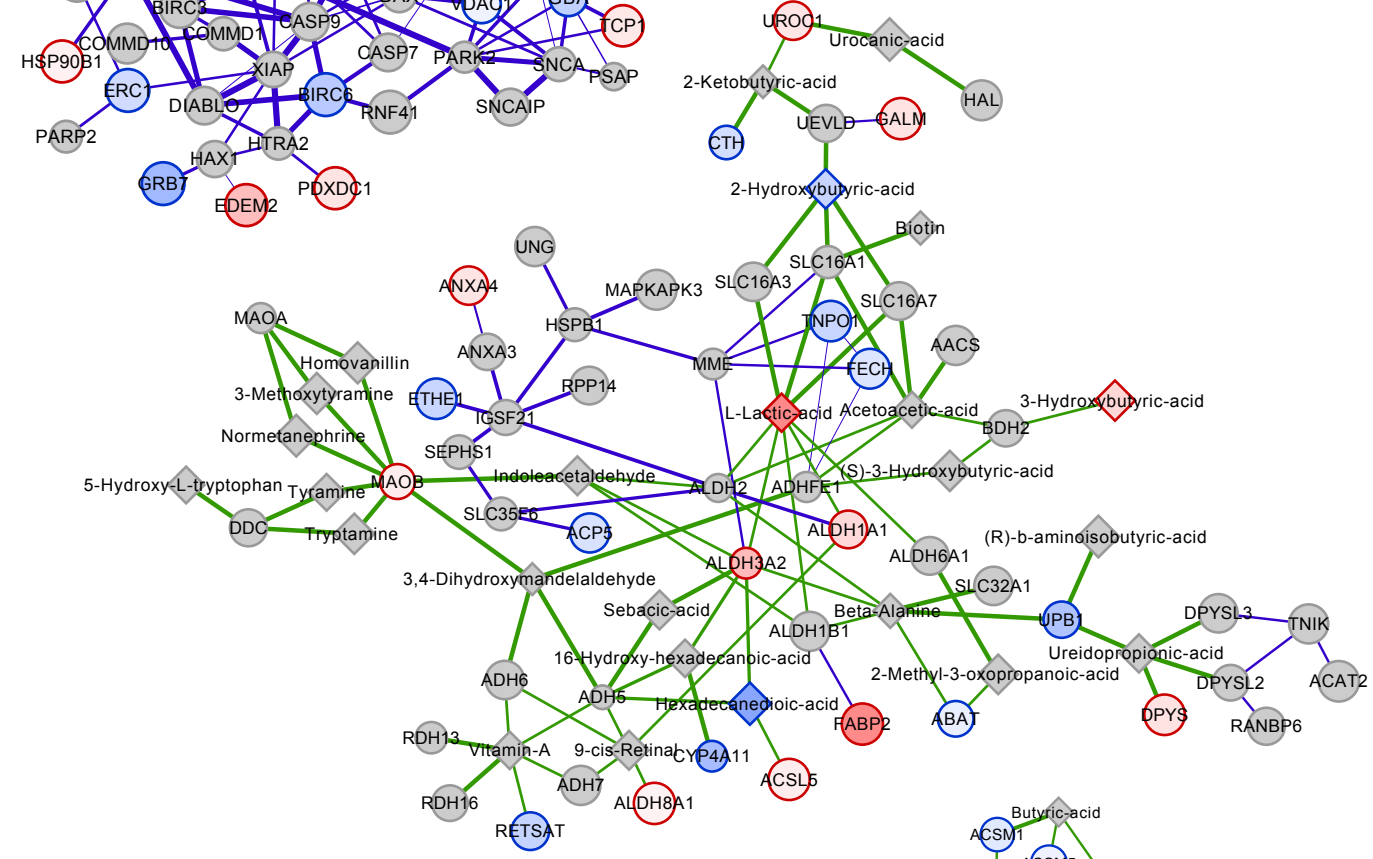
4. Unfolded protein response/ ketone body metabolism



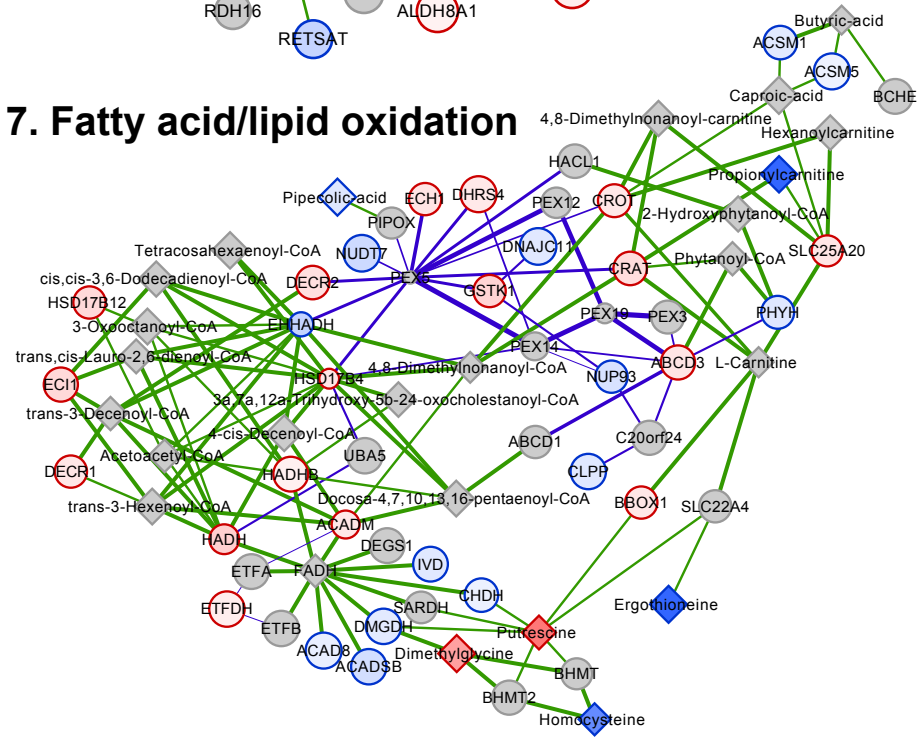
5. Apoptosis



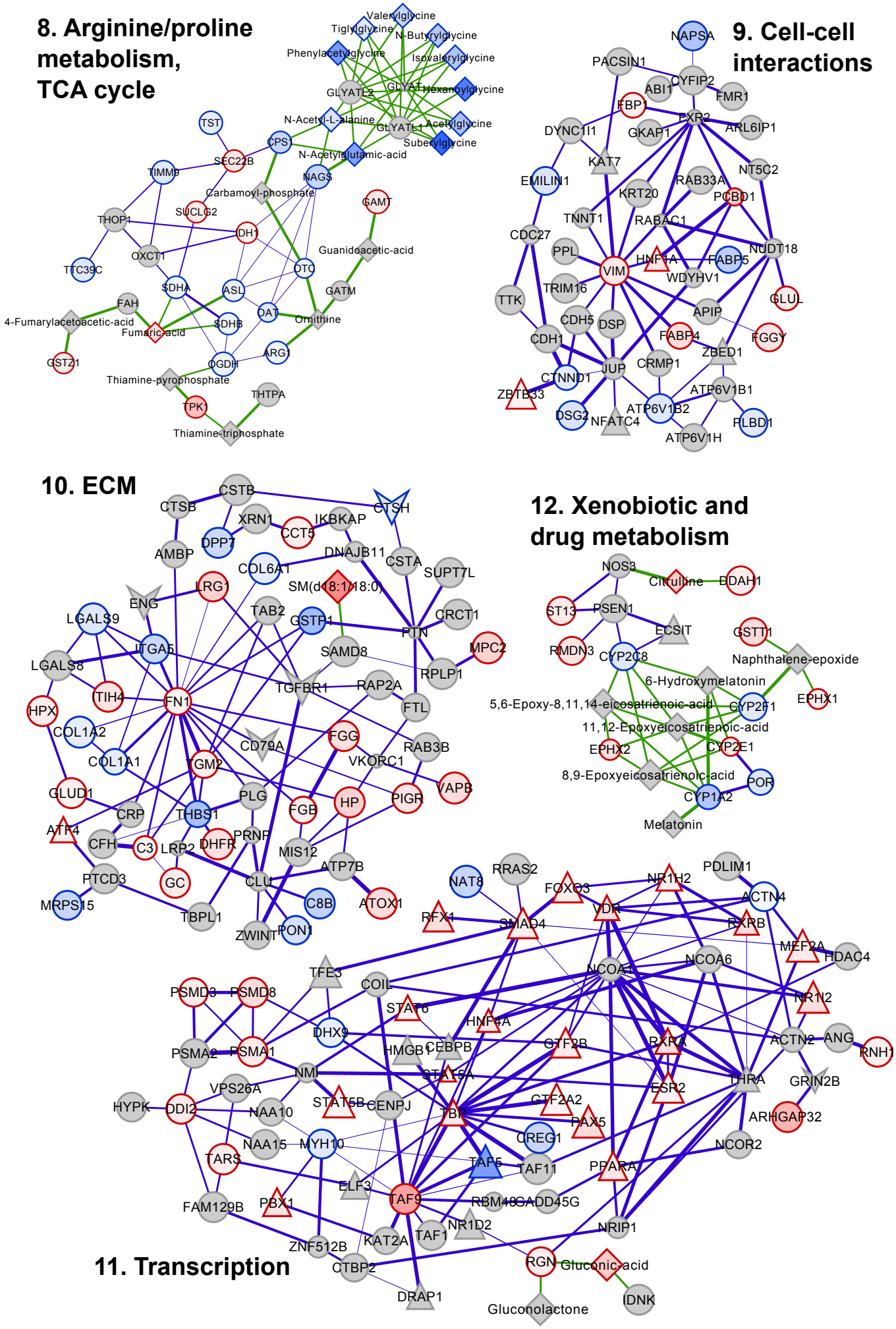
6. Histidine, tyrosine, and retinol metabolism



7. Fatty acid/lipid oxidation

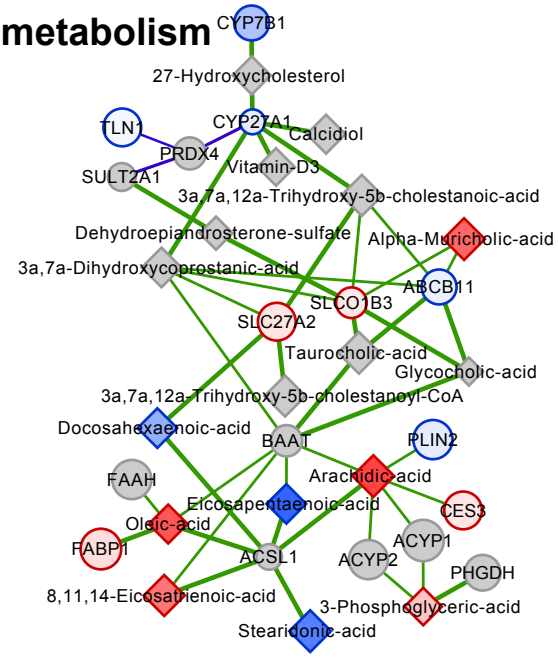


Data S6 (page 3)

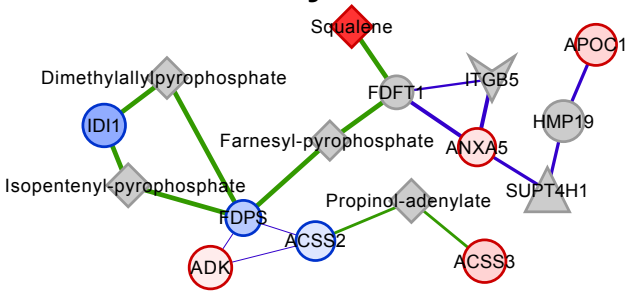


Data S6 (page 4)

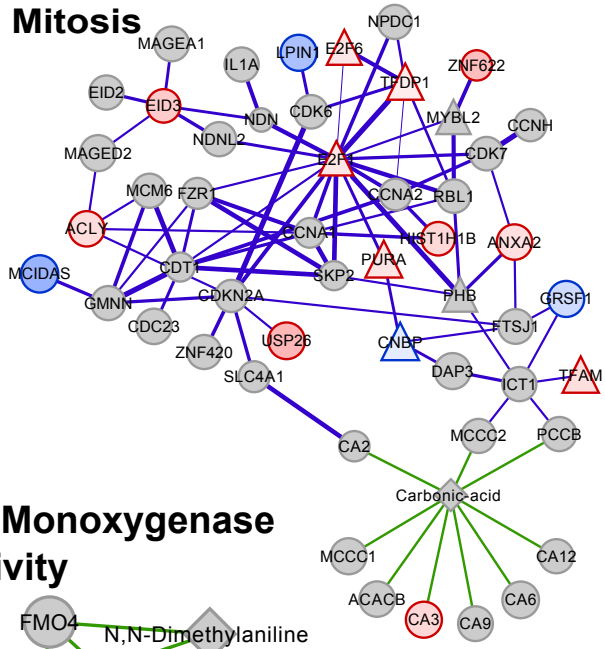
13. Bile acid metabolism



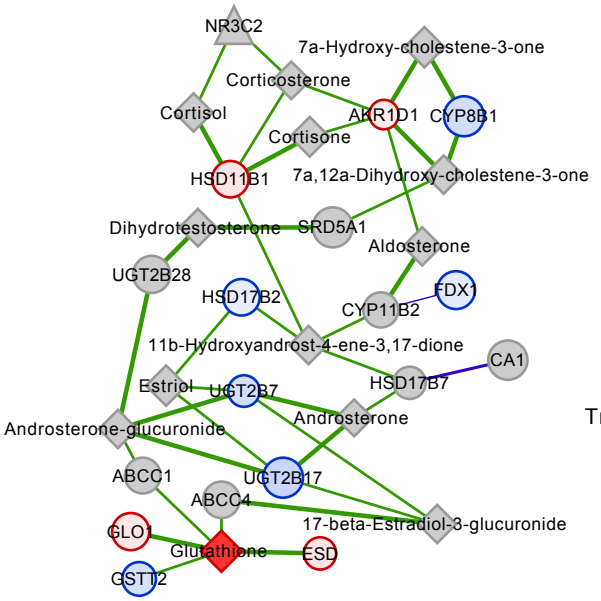
14. Steroid biosynthesis



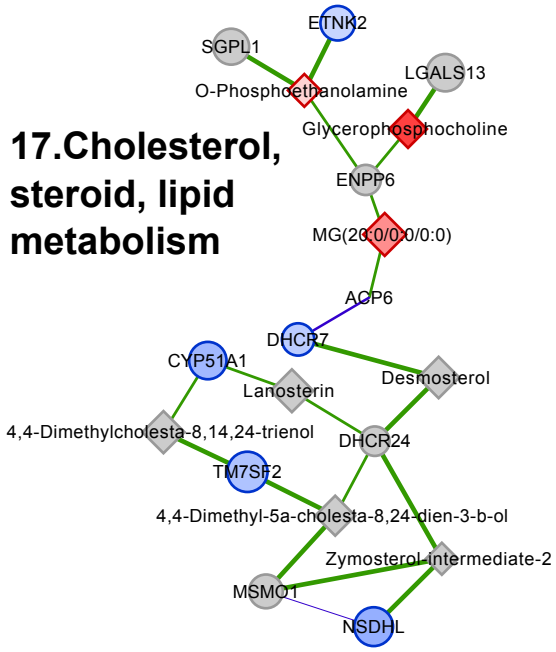
15. Mitosis



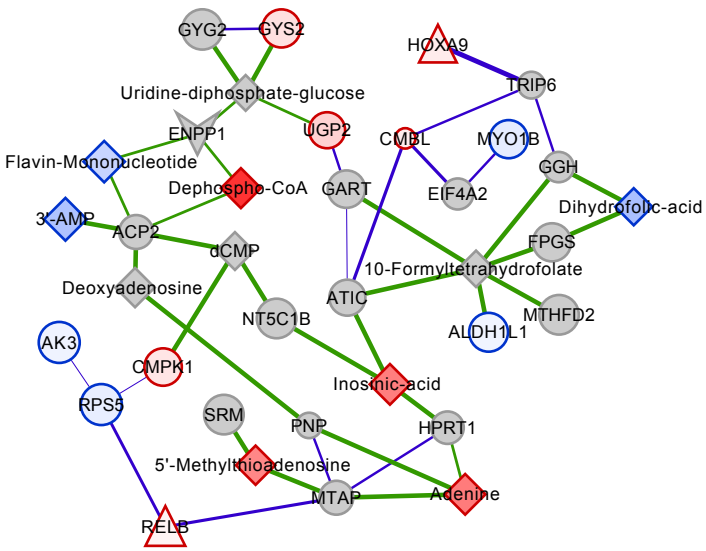
16. Steroid metabolism



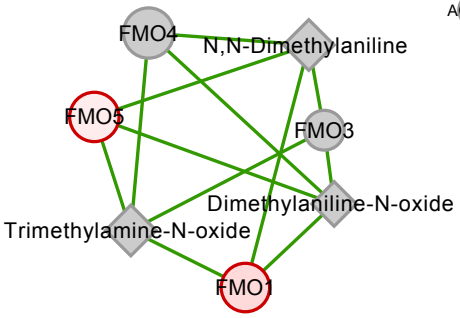
17.Cholesterol, steroid, lipid metabolism



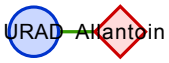
18. Folate, purine/ pyrimidine metabolism



19. Monoxygenase activity



20. N/A



Data S6. PCSF model sub-networks (related to Figures 4 and 5).

Displays for all 20 PCSF model sub-networks, including numeric labels and biological process/pathway enrichments.