Systems biology of diet-induced hepatic insulin resistance

by

Anthony Robert Soltis

B.S. Biomedical Engineering
University of Virginia, 2009

SUBMITTED TO THE DEPARTMENT OF BIOLOGICAL ENGINEERING IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN BIOLOGICAL ENGINEERING
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2017

Signature of Author: _____
Department of Biological Engineering
March 10, 2017

Certified by: _____
Ernest Fraenkel
Professor of Biological Engineering
Thesis Supervisor

Accepted by: _____
Mark Bathe
Associate Professor of Biological Engineering
Chairman, Graduate Program Committee

Thesis Committee:

Douglas A. Lauffenburger
Ford Professor of Biological Engineering, Chemical Engineering, and Biology
Massachusetts Institute of Technology
Thesis Committee Chair

Roger J. Davis
Professor, Program in Molecular Medicine
Howard Hughes Medical Institute
University of Massachusetts Medical School
Thesis Committee Member

Systems biology of diet-induced hepatic insulin resistance

by
Anthony Robert Soltis

Submitted to the Department of Biological Engineering
on March 10, 2017, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Biological Engineering

ABSTRACT

Human obesity is a world-wide health crisis that promotes insulin resistance and type 2 diabetes. Obesity increases intracellular free fatty acid concentrations in peripheral tissues, particularly the liver, which disrupts molecular mechanisms that maintain normal glycemia in response to fasting and feeding. The progression towards outright pathology in response to obesity is a highly complex process that involves coordinated dysregulation of a variety of molecular processes across multiple regulatory levels. The goal of this thesis was to apply a quantitative, multi-omic systems biology approach to the study of obesity-induce hepatic insulin resistance.

We fed male C57BL/6J mice high-fat diets (HFD) to induce obesity and insulin resistance. In the first presented study, our group collected datasets to profile the hepatic epigenomes, transcriptomes, proteomes, and metabolomes of chow diet (CD) control and HFD-fed mice. I extended and applied an established computational modeling algorithm, namely the prize-collecting Steiner forest (PCSF), to simultaneously integrate these molecular data with protein-protein and protein-metabolite interactions into a tractable network model of hepatic dysregulation. This model uncovered a variety of dysregulated pathways and processes, some of which are not well-established aspects of insulin resistance. We further tested and validated some of these model predictions, finding that HFD induces serious architectural defects in the liver and enhances hepatocyte apoptosis.

In the next study, we focused more specifically on hepatic transcription. We fed mice short and long-term HFDs and treated them with the type 2 diabetes drug metformin. Compared to non-treated CD controls, diet exerted the strongest effect on transcription, progressively inducing changes as HFD duration increased. We additionally stimulated mice with insulin and collected temporal transcriptomic profiles. We found that long-term HFD almost completely blunted normal insulin-induced transcriptional changes, but also found a small set of genes that are specifically insulin-responsive in HFD livers. We further characterized one of these genes and provided evidence supporting the notion that aspects of hepatic insulin signaling are intact during insulin resistance.

In another study, we collected transcriptomic and epigenomic data from mice fed a calorie-restricted (CR) diet. Interestingly, we found a small set of genes altered in the same direction by both CR and HFD. We then used chromatin accessibility experiments to infer regulators associated with these gene expression changes and found roles for PPARα and RXRα. We performed ChIP-Seq experiments for these factors and treated mice and primary hepatocytes with a PPARα activator, uncovering a role for PPARα in the regulation of anaerobic glycolysis. We also validated novel predicted target genes of PPARα involved in glucose metabolism.

Finally, we profiled hepatic miRNAs in CD and HFD livers, finding that HFD progressively alters their expression levels. We implemented an enrichment procedure and a network modeling approach to analyze these data. We integrated additional mRNA and epigenetic data to infer miRNAs that may play regulatory roles during insulin resistance. In total, this thesis presents a unique comprehensive approach to the study of diet-induced hepatic insulin resistance that revealed new insights into pathology.

Thesis Supervisor: Ernest Fraenkel
Title: Professor of Biological Engineering

ACKNOWLEDGMENTS

First and foremost, I must thank my thesis advisor, Ernest Fraenkel, for his exceptional guidance and support throughput my time at MIT. I am immensely grateful for the opportunities, resources, and insights he consistently provided without reservation to me and all members of the lab. I truly learned a great deal, was consistently challenged intellectually, and matured as a scientist under his guidance. Thank you. I also must thank the members of my thesis committee, Roger Davis and Doug Lauffenburger. Roger consistently provided critical biological insights that steered and focused my work and Doug helped to frame "big picture" aspects of my project. Their unique approaches and insights enhanced the final output of my thesis work. I also thank Chris Burge for his early service as a member of my thesis committee.

Acknowledgments are due to the many experimental and computational collaborators who worked with me on the major projects presented in this thesis. From the Fraenkel lab: Xiaofeng Xin, Shmulik Motola, Chris Ng, Yoon Yap, Bryan Matthews, and Simona Dalin; from Roger Davis' lab: Norm Kennedy and Santiago Vernia; from Jarrod Marto's lab: Scott Ficarro and Feng Zhou; from Forest White's lab: Nina Lee and Antje Dittmann; from Doug Lauffenburger's lab: Brian Joughin and Jesse Lyons; and from Jason Kim's lab: Randall Friedline. I also thank my undergraduate advisees Minyi Lee, Heather Sweeney, and Yoonjeong Cha: Minyi for her consistent work on mechanistic modeling of networks and Heather and Yoonjeong for their contributions to network modeling of miRNAs.

I also must thank past and present members of the Fraenkel lab for maintaining a hard-working, yet relaxed and good-natured, lab environment: Miriam Adam, Murodzhon Akhmedov, Candace Chouinard, Kimberly DeMayo, Tobias Ehrenberger, Renan Escalante-Chong, Tony Gitter, Sara Gosline, Carol Huang, Amanda Kedaigle, Martina Koeva, Adam Labadorf (the burrito-eating inventor of "the Toro"), Alex Lenail, Jonathan Li, Alice Lo, Denise MacPhail, Pamela Milani, Natasha Patel-Murray, Leila Pirhaji, Gabriela Pregernig, Nurcan Tuncbag, Brook Wassie, Jen Wilson, and Ferah Yildirim. A special shout-out goes to "the Safe Space."

I thank members of the MIT BE class of 2010. In particular: Matt Adendorff and Jorge Valdez who, along with Chris Ng and myself, comprised the league of extraordinary lunch-seeking gentlemen, and Marcus Parrish and Bill Hesse. I also thank other past and present members of the BE community for their friendship and scientific assistance: John Casey, Allison

Claas, Kristi Cook, Jaime Jose, Aaron Meyer, Miles Miller, Chris Pirie, Edgar Sanchez, Jeff Wagner, Joel Wagner, and Annelien Zweemer. In addition, I thank Professors John Essigmann and Steve Hall for allowing my wife and me to live in Simmons Hall for an academic year. I also thank Professor Al Grodzinsky, just for being awesome.

Outside of the lab, I spent a considerable amount of time playing club rugby at MIT. I thank all the past and present members of the club for being all-around great people and for providing a welcome distraction from the rigors of lab work. In particular, I thank Nick Allard for graciously allowing me to live with him in his house in Quincy during the last couple years of my PhD, Alex "Noooh!" Auld, Brendan "Tom Brady" Barschdorf, Zach "the mother hen" Boswell, Mike Byrne, Martin Cleary, Evan Cohick, Keith Forward, "the famous" John Gaffney, Hacksaw, Ognjen "the Ogg-man" Ilic, Jarrad Langley, Brad Lassey, John LeBlanc, James "long-shanks" Long, Ruaridh Macdonald, "Uncle" Mick O'Connor, Frank O'Sullivan, Cody Patrick, James "Patty-cakes" Patterson, "Diamond" Dave Worth, and our esteemed coaches Sean Christie and Brad Dufresne.

I thank Chris Rupert for being a great friend and roommate. Chris was the first person I met when I moved to Boston.

I must especially thank my family. I thank my parents, Bob and Anna Margaret, for their never-ending love and support and for raising three wonderful kids. I thank my brother, Alex, for being my best friend from day ~1000 and on, and his wife Ali. I also thank my younger sister Kayla who is becoming a wonderful young woman.

To my wife, Shannon, I thank you so much for being a loving and compassionate partner. Your love and support has been absolutely transformative to my life, and I look forward to many years together.

Last, but in no way least, I must thank my tiny little dog, Butterstick, a.k.a. Bark Stickington, the B, Stick Monster, the fuzziest one, Butter the B, Butterstick C. Puppydog, etc., etc. She has been a wonderful companion to my wife and me over the years and is a never-ending source of love, joy, and calm.

Let the truth of Love be lighted
Let the love of truth shine clear
Sensibility
Armed with sense and liberty
With the Heart and Mind united
In a single perfect sphere

Neil Peart
VI. *The Sphere (A Kind of Dream)*
Cygnux X-1 Book II: Hemispheres

TABLE OF CONTENTS

LIST OF FIGURES & TABLES

CHAPTER 1

GENERAL INTRODUCTION

1.1. Obesity, insulin resistance, and type 2 diabetes

Human obesity is a major world-wide health crisis. Overweight and obesity are typically crudely defined by the body mass index (BMI), which is calculated as an individual's weight in kilograms divided by his or her height in meters; individuals with BMIs in the range of 25-29.9 are generally considered overweight, while people with BMIs greater than 30 are considered obese. According to the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and the National Health and Nutrition Examination Survey (NHANES), more than two in three adults in the United States are considered overweight or obese, with more than one in three considered obese and one in twenty considered extremely obese (BMI > 40) [1, 2]. Between 2011 and 2014, the Centers for Disease Control and Prevention (CDC) estimated a 36% prevalence of obesity in adults, with a higher prevalence in middle-aged and older adults versus the young and a slightly higher prevalence in women (38.3%) compared to men (34.3%) [3]. According to the World Health Organization (WHO), 600 million people were considered obese in 2014 [4]. The WHO also notes that most of the world's population resides in countries where overweight and obesity kill more people than complications associated with underweight. Compared to normal weight individuals, obesity, particularly higher grade obesity (BMI > 35), is significantly associated with higher overall all-cause mortality rates [5]. Thus, obesity is a highly prevalent condition of significant public health concern across the world.

Obesity is associated with a number of metabolic complications, including metabolic syndrome (characterized by insulin resistance, hyperglycemia, and hypertension), β-cell dysfunction, and, ultimately, type 2 diabetes [6-8]. The CDC projects that the prevalence of type 2 diabetes in the United States will increase to 25-28% by the year 2050 (from 14% in 2010) [9]. Increases in adipose tissue mass as a consequence of obesity enhance the release of free fatty acids (FFAs), proinflammatory cytokines, and hormones (e.g. adiponectin and leptin) from adipocytes and other cell types (e.g. macrophages) [8, 10, 11]. These molecules can trigger inflammation in

adipose and other tissues via, for example, activation of NF-κB and c-Jun NH_2-terminal kinase (JNK) signaling [12]. Released FFAs accumulate in peripheral tissues, including muscle and liver, and modulate normal cellular signaling processes. This leads to insulin resistance in these tissues [13]. Studies in humans have shown that accumulated intracellular FFAs are much stronger determinants of peripheral tissue insulin resistance compared to complications induced by circulating plasma lipids [14].

In healthy individuals, insulin release in response to feeding promotes increased uptake of glucose in peripheral tissues (primarily skeletal muscle) and reduces gluconeogenesis in the liver [15]. The onset of insulin resistance in response to obesity suppresses these mechanisms and promotes hyperglycemia. Obesity and insulin resistance are associated with type 2 diabetes in the presence of pancreatic β-cell dysfunction, which can develop from cellular exhaustion due to enhanced insulin demand, desensitization due to elevated glucose, and reduced cellular mass [16, 17]. The level of β-cell dysfunction is a major distinguishing feature between glucose intolerance (also known as "prediabetes") and pathologic diabetes [17]. First-degree relatives of type 2 diabetes patients, who are at greater risk for developing the disease themselves, show reduced β-cell function even when they are not hyperglycemic [18]. Genome-wide association studies (GWAS) have uncovered many genetic loci linked to type 2 diabetes that, among those that map most confidently to specific genes (e.g. TCF7L2, SLC30A8, CDKN2A, etc.), are involved in Wnt signaling and cell cycle regulation, suggestive of roles that maintain normal β-cell function [19]. Thus, the progression towards type 2 diabetes in response to obesity is a continuous process that can be exacerbated by a variety of individual risk factors.

The majority of pharmacological agents used to treat type 2 diabetes aim to control hyperglycemia [20, 21]. Along with lifestyle alteration and weight loss, metformin, a biguanide that principally acts in the liver to lower hepatic glucose production, is typically prescribed as a first-line treatment due to its safety and low risk of hypoglycemia [22-25]. Gastrointestinal side effects, however, are common in patients taking metformin, particularly in the early stages of treatment. Other classes of type 2 diabetes drugs include sulfonylureas, glinides, α-glucosidase inhibitors, thiazolidinediones (TZDs), insulin, and dipeptidyl peptidase 4 (DPP-4) inhibitors [20, 21, 25]. Sulfonylureas, glinides, and insulin itself enhance insulin secretion to lower blood

glucose, though they can cause hypoglycemia [20, 25]. DPP-4 inhibitors also enhance insulin secretion but do not carry the risk of hypoglycemia. α-glucosidase inhibitors delay carbohydrate digestion and reduce glucose absorption rates, though they also produce gastrointestinal side effects [26]. TZDs are specific ligands for peroxisome proliferator-activated receptor γ (PPARγ) that enhance the sensitivity of muscle, fat, and liver to insulin [27]. Side effects associated with TZDs are weight gain and fluid retention [20]. If glycemic control is not achieved with metformin monotherapy, addition of a second drug is prescribed, with the particular choice dependent on specific patient factors and desired effects [21, 23]. Triple therapy is even recommended in cases where dual therapy no longer controls glycemia. Though many compounds are available and show efficacy in lowering blood glucose concentrations, cost, side effects, and patient-specific contraindications and inefficacies are still problematic; thus, novel therapeutics for the treatment of type 2 diabetes are still needed.

1.2. Mechanisms of hepatic insulin resistance

The work presented in this document principally focused on the effects of obesity in the liver. The liver is an insulin-sensitive organ that is critical for the maintenance of normal glucose homeostasis and overall metabolic health. Fasting hyperglycemia, a critical feature of insulin resistance and type 2 diabetes, primarily results from the inability of the liver to properly shut down hepatic glucose production in response to insulin [15, 28, 29]. Over-nutrition leading to obesity (e.g. by consumption of a high-fat diet) prevents insulin-mediated inhibition of hepatic glucose production [30]. The critical importance of the liver in these contexts is particularly highlighted by the widespread use of metformin to treat type 2 diabetes, which inhibits hepatic glucose production and reduces plasma triglyceride levels [24]. The mechanism(s) by which these effects are accomplished by metformin are not fully understood, though possibilities include disruption of gluconeogenic enzyme transcription by AMPK-dependent CREB binding protein (CBP) phosphorylation [31].

Insulin action in the liver is generally understood through the lens of the canonical insulin signaling pathway (**Figure 1-1**) [29, 32, 33]. This pathway is initiated when insulin binds to and activates the insulin receptor, which is itself a tyrosine kinase. The receptor phosphorylates

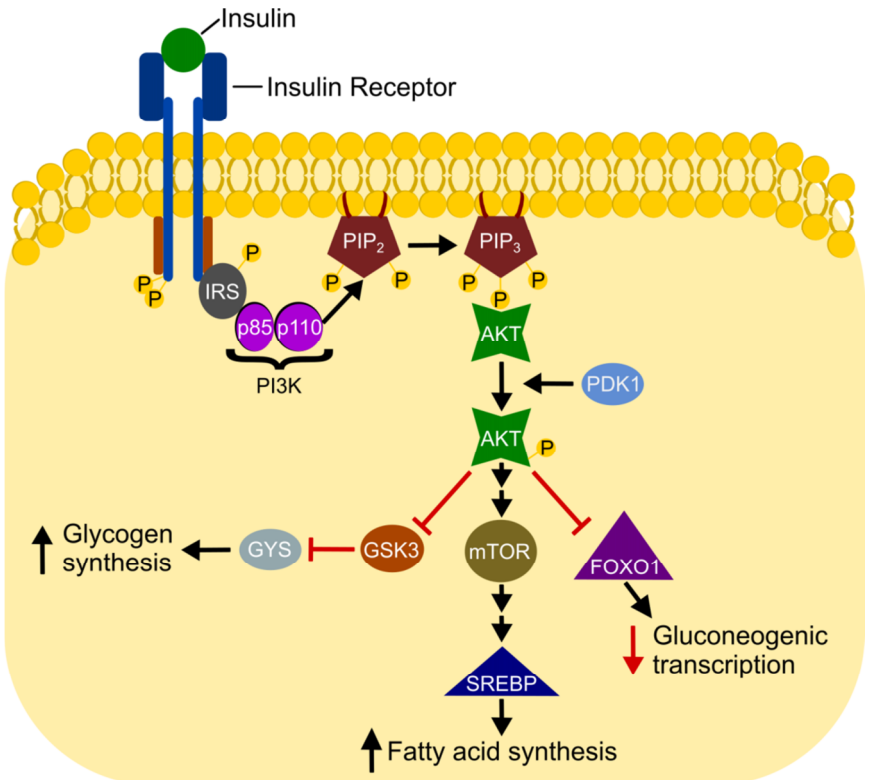


Figure 1-1. Canonical insulin signaling pathway. See text for details and explanation. Abbreviations: insulin receptor substrates (IRS); phosphatidylinositol-3-OH kinase (PI3K); phosphatidylinositol (4, 5) bisphosphate (PIP_2); phosphatidylinositol (3, 4, 5) triphosphate (PIP_3); protein kinase B (AKT); 3-phosphoinositide dependent protein kinase-1 (PDK1); glycogen synthase kinase 3 (GSK3); glycogen synthase (GYS); mechanistic target of rapamycin (mTOR); sterol regulatory element binding protein (SREBP); forkhead box protein O1 (FOXO1).

insulin receptor substrate (IRS) proteins, particularly IRS1 and IRS2 in the liver, and these bind phosphatidylinositol-3-OH kinase (PI3K). The catalytic subunit of PI3K (p110) phosphorylates phosphatidylinositol (4, 5) bisphosphate (PtdIns(4,5)P_2 or PIP_2) to produce PIP_3, which recruits the downstream effector AKT to the plasma membrane. 3-phosphoinositide dependent protein kinase-1 (PDK1) then phosphorylates and activates AKT. Activated AKT translocates to the cytoplasm where it phosphorylates and inactivates glycogen synthase kinase 3 (GSK3) to activate glycogen synthase, thereby promoting storage of glucose as glycogen. AKT also phosphorylates and inactivates by nuclear exclusion forkhead box protein O1 (FOXO1), which decreases transcription of genes encoding gluconeogenic enzymes.

This regulatory pathway is dysregulated in response to fatty acid accumulation in the liver. In rats, insulin resistance can be induced by dietary fat feeding independent of obesity [34]. This is a specific consequence of hepatic insulin resistance. As early as three days after the start of high-fat feeding, liver triglyceride and fatty acyl-CoA content increases and inhibition of endogenous glucose production by insulin is diminished [35]. This occurs without significant increases in muscle fat content. These changes are accompanied molecularly by hepatic decreases in IRS1 and IRS2 tyrosine phosphorylation, decreased IRS1-PI3K and IRS2-PI3K activity, decreased AKT activity, and reduced inhibition of GSK3 activity. These results may reflect a signaling switch from insulin-stimulated tyrosine phosphorylation cascades to fatty acid-induced serine/threonine regulation [8]. To this, hepatic JNK activity is increased in mice in response to diet-induced obesity; ablation of JNK1 improves insulin sensitivity and insulin receptor signaling [36]. JNK has been shown to associate with IRS1 and phosphorylate serine 307, leading to inhibition of insulin-stimulated tyrosine phosphorylation of this target [37]. Hepatic lipid accumulation can also lead to protein kinase C epsilon (PKCε) activation, which associates with the insulin receptor and impairs its kinase activities [38]. Knock-down of PKCε in the livers of rats by antisense oligonucleotides protects them from lipid-induced insulin resistance [38].

This canonical view of hepatic insulin signaling and its relationship to type 2 diabetes pathogenesis has been challenged by more recent findings. Diabetic mice experience "selective insulin resistance," whereby insulin fails to suppress hepatic glucose production but is still able to induce lipogenesis via SREBP-1c production, indicating only partially dysregulated hepatic insulin signaling in the disease state [39, 40]. Liver-specific insulin receptor knock-out (LIRKO) mice are totally insulin resistant and show severe glucose intolerance and elevated fasting insulin levels, along with an inability to suppress hepatic gluconeogenesis, but actually show slight fasting hypoglycemia as age increases [41, 42]. Whole-body knock-out of the insulin receptor in mice, however, does cause postnatal diabetes and early death due to ketoacidosis [43]. More acute suppression of the hepatic insulin receptor by antisense oligonucleotides in C57BL/6J mice, resulting in ~95% knock-down of hepatic protein expression of this target, impairs downstream insulin signaling in the liver but does not alter rates of glucose production [44]. These traditional mechanisms of glucose regulation are particularly complicated by the observation that mice lacking *Akt1*, *Akt2*, and *Foxo1* by triple knock-out are still able to suppress

hepatic gluconeogenesis normally in response to insulin [45]. Double liver insulin receptor and *Foxo1* (L-IRFoxo1DKO) knock-out mice also are glucose tolerant and able to normally suppress hepatic glucose production and gluconeogenic gene expression in response to insulin, processes which are disrupted with LIRKO alone [46]. Additional recent evidence suggests that liver insulin signaling may mostly be intact during type 2 diabetes, while increased intrahepatic FFAs (e.g. acetyl CoA) due to enhanced lipolysis from white adipose tissue inhibit suppression of glucose production by mechanisms independent of the canonical insulin signaling [47, 48].

Years of experimental analysis have clearly demonstrated the critical role of the liver in the pathogenesis of type 2 diabetes, though recent findings highlight that our understanding of the mechanistic underpinnings of its role are incomplete. Therefore, new studies are needed to interrogate novel hepatic mechanisms that promote and maintain metabolic disease. Such findings may expand the scope of therapeutic strategies to treat type 2 diabetes.

1.3. Experimental models of hepatic insulin resistance

Crucial to the study of any complex human disease is the selection of an appropriate experimental model that maximally reproduces aspects of the disease consistent with human pathogenesis and pathology, while simultaneously allowing for proper experimental control, reproducibility, and hypothesis testing. A variety of models are available for such purposes, each possessing their own benefits and limitations. For a given scientific question, the use of multiple model types can help confirm observations made from initial screens in a chosen system. Here, I review three major types of experimental models applied to the study of insulin resistance and type 2 diabetes: cellular models, genetically modified rodent models, and diet-induced obese rodent models.

Cellular models: A variety of primary cell culture and immortalized cell line systems are available for the study of insulin resistance and related complications [49]. Primary hepatocytes, along with other hepatic cell types (Kupffer, stellate, etc. cells), can be cultured from human and rodent samples and grown in either mono or co-culture systems [49, 50]. To mimic obesity-induced insulin resistance in such systems *in vitro*, cells can be cultured in the presence of

monounsaturated (e.g. oleate) and saturated (palmitate) fatty acids to enhance intracellular FFA concentrations [50, 51]. Additional treatments are available to induce insulin resistance in cellular models, including TNFα, hypoxia, high concentration insulin, and dexamethasone [52]. Outside of primary cell cultures, the common perpetual human liver cancer line HepG2 is also used in these contexts and also often cultured in the presence of FFAs to mimic obesity-induced insulin resistance [53].

A variety of complications are associated with such *in vitro* systems. The availability of human material for such systems is limited, and cultures derived from rodents can de-differentiate and lose tissue-specific functions as culture time increases [49]. Cell lines like HepG2 do not suffer from these same issues; however, gene expression microarray studies have shown extensive differences between HepG2 expression profiles and those of human primary hepatocytes and frozen liver tissue [54]. Here, expression profiles between primary hepatocytes and liver tissue show greater similarity than comparisons involving HepG2, limiting the translatability of findings derived from such models. Still, comparisons of expression profiles between mono culture primary human hepatocytes and those of intact liver tissue also display considerable levels of gene expression differences (~22% of genes in some studies) [54]. These results demonstrate clear trade-offs associated with ease of use, availability, and down-stream relevance to the actual human condition when using such cellular models.

Genetic rodent models: Several genetically modified rodent models are available and commonly used for the study of obesity and diabetes. *Ob/ob* mice (where *ob* stands for "obese") were discovered in 1949 and characterized in 1994 to possess a frame shift mutation in the leptin gene that creates a premature stop codon [55]. Leptin protein is important for appetite control; mice possessing this mutation are hyperphagic and exhibit obesity, insulin resistance, and type 2 diabetes. *Db/db* mice (*db* for "diabetes"), discovered in 1966, possess a G-to-T point mutation in the leptin receptor gene [55]. This mutation impairs leptin signaling and induces obesity, insulin resistance, and eventual hyperglycemia. Apolipoprotein E3-Leiden (ApoE3L) mice are transgenic animals crossed to the C57BL/6J background that express the human APOE*3Leiden and apoC1 gene cluster. These mice display a lipoprotein profile that closely resembles humans [56]. Additional genetic rodent models, some polygeninc, include Zucker fatty rats (*fa/fa*), the

New Zealand obese (NZO) mouse, the Tsumura Suzuki obese diabetes (TSOD) mouse, and the melanocortin 4 receptor (MC4R) disruption mouse [49, 55, 57]

While genetically modified rodents are widely used in diabetes research, the specific mutations and/or modifications used to induce these complications do not always reflect aspects of the human disease [58]. For example, the *ob* mutation is rare in humans and leptin is not strongly associated with human type 2 diabetes [49, 59]. Mutations in rodents can therefore induce additional dysregulation of pathways and processes that are dissociated from the human condition, complicating the translation of findings from these animals.

Diet-induced obese rodent models: Diets containing excessive nutritional content, especially high-fat diets (HFD, 30-75% total calories from fat), are commonly used to induce experimental obesity [49, 55]. The HFD-fed C57BL/6J mouse, which was first described in 1988, is the most commonly used diet-induced obese rodent model [57, 60, 61]. Critically, C57BL/6J mice fed a HFD develop complications consistent with human metabolic syndrome, including obesity, hyperinsulinemia, hypertension, hyperglycemia, and insulin resistance [62]. The effects of high-fat feeding in mice are strain dependent [55]. For instance, C57BL/KsJ mice display weaker phenotypic characteristics compared to C57BL/6J mice when fed a HFD. While diet-induced obese models more faithfully reproduce human complications on the road to pathology, limitations associated with differences between rodents and humans still must be considered when interpreting results obtained from such studies.

Considerations regarding reproducibility, availability, and relevance to the human condition are critical when choosing an experimental model for the study of any disease. While all models have limitations, critical insights relevant to human disease can be obtained from well-designed studies. In the work presented in this thesis, our research team utilized a HFD-fed C57BL/6J mouse model to study obesity-induced hepatic insulin resistance. This model choice is highly appropriate given its reproducibility and established consistency with human progression towards metabolic syndrome and type 2 diabetes.

1.4. Systems biology

Biological processes that promote and maintain disease are often highly complex and involve molecular mechanisms operating across many levels of biological regulation. In the context of insulin resistance and type 2 diabetes, we now appreciate that abnormalities occur across many such levels [63]. For example, GWA studies have shown that genetic variations in genes encoding transcription factors (e.g. *TCF7L2* and *PPARG*), ion channels (*KCNJ11*), and insulin signaling pathway members (*IRS1*) are strongly associated with human risk for type 2 diabetes [64]. Thus, studies that aim to reveal the complex mechanisms associated with such diseases may reveal new avenues for therapeutic intervention.

The field of systems biology seeks a global understanding of disease by embracing these inherent complexities within true biological systems [65]. Systems biology approaches draw upon the wealth of information generated from molecular biology and biochemistry in the past to engineer systems tools (e.g. computational models) that not only explain observed biological phenomena, but are capable of predicting responses to new perturbations, thereby uncovering non-intuitive, emergent system properties that are only observed when biological components are studied in context with one another [66]. Systems biology as a paradigm is quantitative in its descriptions of physical interactions and in its data collection methods and philosophically departs from past "reductionist" approaches to biology in favor of more "holistic" methods [67]. This conceptual framework is elegantly described by the following quote from Hiroaki Kitano [65]:

"*Identifying all the genes and proteins in an organism is like listing all the parts of an airplane. While such a list provides a catalog of the individual components, by itself it is not sufficient to understand the complexity underlying the engineered object*."

Systems theory applied to biology initiated in the early to mid-twentieth century. The Austrian biologist Ludwig von Bertalanffy was a key architect of the general systems theory, which stressed holistic understanding of systems over reductionism [68]. In 1952, Alan Hodgkin and Andrew Huxley published a mathematical model describing the ionic mechanisms that generate, propagate, and terminate action potentials in the squid giant axon using systems of nonlinear, ordinary differential equations (ODEs) [69]. For their work, they, along with John Eccles, were

awarded the 1963 Nobel Prize in Physiology or Medicine. These methods were subsequently (and still are) applied to related cellular phenomena, notably the cardiac action and pacemaker potentials, which were first modeled in the 1960s [70] and expanded upon in subsequent decades [71, 72]. Since the early 2000s, following completion of the human genome project and the advent of new high-throughput "omic" technologies, systems biology has widely been applied to the study of diverse biological problems and is beginning to mature as a field [65, 73, 74]. Systems biology studies can vary in size and scope, from <10 component signaling pathway models [75] to genome-scale reconstructions of human central metabolism [76] or whole-cell models of bacterial pathogens [77], and can employ a wide variety of experimental and computational methods.

The application of a systems biology approach to the study of insulin resistance may reveal novel insights into molecular mechanisms that drive pathology and highlight entry points for therapeutic intervention [78]. The work presented in this thesis utilized such an approach toward these goals.

1.5. "Omic" datasets for system biology

A branch of modern systems biology aims to collect as much relevant data as possible from as many biological regulatory levels as can be confidently measured to achieve this goal of holistic characterization of a given system. To this end, "omic" data technologies have emerged as valuable tools for systems biology [79]. The Oxford English Dictionary defines the use of the suffix *–ome* in the context of cell and molecular biology as "all of the specified constituents of the cell, considered collectively or in total." In keeping with this definition, modern omic methods aim to comprehensively profile a given biological "ome." For example, transcriptomic studies aim to capture and quantify all cellular RNA transcripts, proteomic studies measure total or modified protein levels, etc. Many omic techniques collect this information in an unbiased fashion, though variants operate in targeted manners to profile specific sub-sets of molecules. A variety of experimental methods are used extensively today to collect omic data.

Completion of the human genome project inspired rapid technological growth that has produced powerful strategies amenable to high-throughput analysis of the genome, epigenome, and transcriptome. These technologies allow for the measurement of hundreds of millions to billions of short DNA sequence fragments, exploring a wide array of biological phenomena in a single experimental run [80]. Such techniques, often appended with the suffix "-Seq" for "sequencing," allow for comprehensive analysis of genomic variants, transcriptional output (e.g. mRNAs, micro RNAs), protein-DNA interactions, epigenetic modifications (e.g. DNA methylation, histone modifications), and chromatin accessibility on a genome-wide scale. As these technologies improve and as costs decline, their uses in the modern biological sciences are rapidly increasing.

RNA-Seq is a commonly used high-throughput sequencing methodology. It is a highly-reproducible transcriptome-wide approach that can rapidly quantify all or subsets of the RNA species in a cellular system [81, 82]. This approach has distinctive technical advantages in terms of sensitivity over earlier transcriptomic methods, such as microarrays [83]. RNA-Seq protocols are used to profile mature mRNAs (mRNA-Seq), small RNAs (smRNA-Seq, e.g. for micro RNAs), or total RNA populations following ribosomal RNA depletion (Ribo-Zero RNA-Seq) [82]. These methods can be used to quantify and compare gene expression levels, individual transcript levels, and alternative splicing [84-86].

High-throughput sequencing protocols are also used to profile the physical genome [87] and epigenome [88]. Whole-genome sequencing can provide critical information regarding alterations physically contained within the DNA sequence, including base mutations, insertions, deletions, inversions, etc. In contrast, study of the epigenome provides insight into how the physical genome is regulated and can provide complementary information to gene expression studies. The arena of epigenomics encompasses covalent histone modifications, DNA accessibility, and DNA methylation [89]. Chemical modifications to histone proteins induce electrostatic changes that alter the binding affinities of histone proteins for DNA, thereby creating "active" or "repressed" regions throughout the genome that regulate gene expression. Examples of such modifications include tri-methylation of lysine 4 on histone protein H3 (H3K4me3), which "marks" active promoters of expressed genes [90], mono-methylation of

lysine 4 on H3 (H3K4me1), which marks active and poised promoters and distal enhancers [91], acetylation of lysine 27 on H3 (H3K27Ac), which marks active promoters and enhancers [91], and tri-methylation of lysine 27 on H3 (H3K27me3), which is associated with gene repression [92]. Specific combinations of these marks influence functional consequences related to gene transcription; these combinatorial effects define what is known as the "histone code" [93]. The most common high-throughput method used to profile histone modifications is chromatin immunoprecipitation followed by sequencing (ChIP-Seq), which uses antibodies targeting specific epitopes on these modified proteins to isolate bound chromatin fragments for subsequent sequencing [94]. ChIP-Seq is also widely used to directly profile protein-DNA interactions of specific regulatory proteins, generally transcription factors. DNA accessibility assays, or "open" chromatin studies, indirectly profile the epigenome in the sense that they do not target a specific physical modification; rather, they provide information on the state of the epigenome and profile active regulatory regions across the genome. DNA accessibility profiling methods include DNase-Seq [95], formaldehyde-assisted isolation of regulatory elements (FAIRE)-Seq [96], and the assay for transposase-accessible chromatin (ATAC)-Seq [97], all of which provide similar information via different experimental methods. Finally, DNA methylation is a chemical modification that occurs on the bases of the DNA molecule itself, particularly cytosine. Sequencing-based whole-genome and reduced-representation protocols are available to profile this epigenomic modification [98, 99].

Proteomics is the study of all the expressed proteins in a biological sample. Such studies focus on total protein levels or chemically modified species, e.g. via phosphorylation, ubiquitination, acetylation, etc. [100]. Most high-throughput proteomic studies are conducted using mass spectrometry-based methods that specifically quantify peptides derived from protease-cleaved full proteins. Traditionally, these studies are run with untargeted "shotgun" protocols, though targeted approaches, including multiple reaction monitoring, target peptide monitoring, and data-dependent acquisition, are increasingly being used to quantify specific molecules of interest [101, 102]. Both relative and absolute quantification methods are used in proteomic studies, though the former is most common. Relative quantification methods include stable-isotope labeling with chemical tags (e.g. isobaric tags for relative and absolute quantification [iTRAQ]) [103], label-free quantification [104], and *in vivo* metabolic stable-isotope labeling (e.g. stable-

isotope labeling by amino acids in cell culture [SILAC]) [105]. Such methods are used to measure biochemical properties of proteins (e.g. degradation rates) [106] and are continuously applied toward the goal of profiling the full human proteome [107]. Proteomic methods are not only applicable for the identification and quantification of individual proteins and associated rate parameters, but also for the interrogation of protein-protein interactions [100]. Interaction information can be derived from mass spectrometry-based approaches by several methods, including immunoprecipitation [108] and proximity labeling [109]. Known protein-protein interactions currently number in the hundreds of thousands [110]; this information is critical for systems biology as it allows for interpretation of high-throughput proteomics (and other omic) datasets in the context of networks of physical interactions.

Metabolomics is the collective study of all the small molecule species within a biological sample. Metabolomics is now being used to identify new mechanisms and biomarkers of disease [111]. Application areas include prostate cancer [112], glioblastoma [113], Crohn's disease [114], and type 2 diabetes [115, 116]. Metabolites are the essential constituents of all the biochemical species that coordinate cellular activities; the full compendium of metabolites present in the human metabolome consists of ~5,000 known small molecules [117]. Today, several targeted and non-targeted experimental strategies are capable of detecting and quantifying hundreds to thousands of small molecules in a sample of interest [111, 118]. These methods typically use gas chromatography and/or electrospray ionization (in positive and/or negative modes) with one or two rounds of mass spectrometry to identify and quantify small molecules. These methods are of critical importance for the study of disease, particularly in the context of the metabolic conditions of focus in this thesis.

A final omic regulatory level that is increasingly being analyzed in the context of human disease is the microbiome, which is comprised of the small microbial communities on and within the body. Dysregulation of the microbiome (termed "dysbiosis") can affect metabolism and drug interactions and may be associated with metabolic diseases like type 2 diabetes [119, 120]. For instance, changes in mouse gut microbiota have been shown to affect feedback loops between the liver and gut that control bile acid metabolism [121]. Studies of the microbiome typically involve the identification of microbial species by sequencing of phylogenetically informative

markers, such as the ribosomal small subunit gene [119]. The microbiome thus represents another regulatory layer with potential application to systems biology studies of metabolic diseases.

1.6. Bioinformatics, statistics, and computational modeling for systems biology studies

Statistical methods and bioinformatics are absolutely essential to omic systems biology studies. Statistical methods identify significant signals in omic datasets (e.g. differentially expressed genes between conditions, enriched ChIP signal intensities over background noise). Analysis of omic datasets can require both continuous (e.g. χ^2-distribution) and discrete (Poisson, negative binomial) probability distributions for accurate modeling and inference. Such choices are dependent on the type of data and the methods of quantification. Bioinformatics as a field combines statistics, computer science, mathematics, and engineering to address biologically-related problems. Applications include cataloging of biological information, including genomic sequences and protein interactions, sequence assembly, and methodological development for the analysis of genomic sequence elements. The University of California Santa Cruz (UCSC) Genome Browser provides such genomic information, including reference genomic sequences for humans and model organisms, along with analytical tools (e.g. fast sub-sequence extraction from large genomes) [122]. Additional examples of Bioinformatics applications include tools for the identification of transcription factor binding sites using Bayesian modeling of epigenetic features (e.g. chromatin accessibility) and sequence information (e.g. known DNA recognition sequences for factors or "motifs") [123, 124] and for *de novo* (MEME, HOMER) or hypothesis-based (THEME) discovery of sequence motifs in sets of relevant short DNA sequences [125-127]. Catalogues of experimentally determined and predicted DNA sequence motifs for regulatory factors are available from multiple data sources, including JASPAR [128] and TRANSFAC [129], and can aid analysis of genomic features identified by omic datasets. Protein-protein interaction information is also readily available from data sources, including iRefIndex [110].

Analysis of individual datasets is critical for interpretation of omic information; however, modern systems biology studies now collect multiple omic datasets from the same or related

samples towards the goal of obtaining holistic systems views. Such multi-omic studies require analytical techniques that jointly interrogate complimentary biological information. Methodological development to this end is a current and emerging field in the computational sciences [130, 131]. Bersanelli *et al*. (2016) distinguish between two major classes of integrative methods: network-free and network-based [130]. The former refers to methods that do not utilize prior assumptions or knowledge of relationships between biological species (e.g. protein-protein interactions), while the latter utilizes such data. Network-free methods typically employ some form of correlation or regression analysis to establish relationships between predictor (e.g. gene and protein expression levels) and response (e.g. measures of phenotype) variables. Such methodologies include simple pair-wise correlation network analyses [132, 133], though more sophisticated multivariate statistical routines are also commonly used. Partial least squares regression (PLSR) is a multivariate statistical procedure that attempts to find a multidimensional representation of predictor variables that best explains (i.e. maximizes the covariance between) response variables, and is especially suitable when using large numbers of multi-collinear predictor variables [134]. Methods implementing variations of PLSR include stochastic multivariate regression [135], sparse PLS (used in the Integromics package) [136], multi-block PLS [137], and orthogonal PLS [138]. Additional related multivariate statistical procedures include independent component analysis [133], extended canonical variate analysis [139], and principal component regression [140]. Sample discrimination and/or phenotype prediction are generally improved with these methods when multiple datasets are considered. Bayesian network-free methods are also available for multi-omic data analysis, including iCluster [141], which jointly clusters multiple omics datasets, and multiple dataset integration (MDI) [142], which accomplishes the same clustering task as iClulster, but via Dirichlet-multinomial allocation mixture modeling. These network-free methods are indeed able to uncover relationships between multiple types of data, either through supervised or unsupervised approaches, though they provide little information as to *how* or *why* specific signals (e.g. highly weighted proteins in a PLSR model) contribute to observed responses.

Network or pathway-based computational modeling methods directly utilize information regarding the underlying mechanisms by which molecules drive biological functions. This includes static and dynamic interaction information. Modeling is highly critical for systems

biology as our intuition of systems-level behavior diminishes as the number of components and potential connections between them increases. A wide array of modeling approaches, varying in scope, complexity, and abstraction, exist for systems analyses. ODE models have been used to study a wide-range of biological pathways and processes, including *E. coli* chemotaxis [143], apoptosis [144], cardiac signaling and electrophysiology [145, 146], and EGFR signaling [147]. These models attempt to simulate the temporal and spatial dynamics of a biological system using as much detailed mechanistic information as possible. This typically requires the knowledge and/or estimation of hundreds of free parameters. As model size increases, ODE modeling is complicated by such parametric requirements; however, studies of systems biology models have revealed that many parameter regimes are "sloppy," meaning that many parameters can vary over wide ranges without affecting the overall system's behavior, and that many parameters can be simply approximated by order-of-magnitude estimates without altering overall model performance [148, 149]. More abstract formulations, such as logic-based approaches (e.g. Boolean [150] and fuzzy logic [151]), require less parameter information. These approaches, however, possess limited ability to accurately model temporal dynamics and absolute concentrations of biological species. Still, some modeling frameworks abstract parameter information away completely. This is often done in the field of metabolic network analysis with approaches like "flux balance analysis" (FBA) [152] or "flux variability analysis" (FVA) [153]. Genome-scale models of central metabolic pathways typically use a "stoichiometric matrix" $\boldsymbol{S}$ to describe the system, where the rows of $\boldsymbol{S}$ are metabolites, the columns are metabolic reactions, and the elements are reaction stoichiometric coefficients. FBA and FVA use constraint-based linear programming to solve the matrix equation $\boldsymbol{S \cdot v = 0}$, where $\boldsymbol{v}$ is a matrix of steady-state "fluxes" through individual reaction paths, to optimize some biological feature (e.g. growth). The steady-state assumption removes the parametric burden and allows for analysis of normal and perturbed metabolic flux spaces through large networks. Yizhak *et al.* devised the integrative omics-metabolic analysis (IOMA) methodology, which combines proteomic and metabolomic data with such genome-scale metabolic models [154]. IOMA uses quadratic programming to solve for the steady-state fluxes through individual metabolic reactions, constraining fluxes to match estimated values from measured enzyme and metabolite concentrations.

A critical issue associated with the pathway-based modeling methods reviewed thus far is the *a priori* requirement of the specific underlying network structure. Often omic datasets measure species that are not part of well-defined canonical signaling pathways. Restricting systems analyses to only such molecules can prevent the discovery of novel aspects of disease [155]. These issues can be directly addressed with network modeling methods. Such methods reveal underlying network structures from experimental data of physical interactions [156]. Naïve methods, however, can often create uninformative "hairball" networks that are of little use for subsequent analyses [157]. More tractable network modeling approaches have been developed to address these issues, including methods that solve the prize-collecting Steiner tree problem [155, 158-160], that perform minimum-cost flow optimization [161, 162], or that model network diffusion [163]. Network modeling methods are thus highly amenable to multi-omic data integration.

1.7. Prior omic systems biology studies of obesity, insulin resistance, and type 2 diabetes

To date, several studies have attempted to apply omic systems biology approaches to the study of obesity, insulin resistance, and type 2 diabetes, though the diversity of omic levels profiled, the breadth of data integration, and the application of computational modeling differs considerably. Most studies have generally focused on a single type of omic data, particularly transcriptomics [63]. Transcriptomic studies have analyzed responses to various dietary interventions (e.g. high-fat, high-carbohydrate) in various tissues (liver, fat, muscle) across a number of experimental models. In general, the biological processes altered by these diets are consistent across models and treatment conditions. As an example, Radonjic *et al.* (2009) collected transcriptomic data from chow and HFD-fed ApoE3Leiden mice at various time points following the start of HFD using microarrays and noted a switch from early (first few days and weeks) inflammatory profiles to late (8-16 weeks) steatotic expression patterns, the latter resulting from up-regulation of genes associated with lipogenesis, lipid accumulation, and fatty acid synthesis [56]. This same group also profiled hepatic transcriptional responses to various lifestyle and chemical treatments following HFD in *Ldlr*$^{-/-}$ mice [164]. Kelder *et al.* (2011) also collected hepatic transcriptional data from low- and high-fat diet-fed mice following glucose treatment and mapped differential

genes onto known pathways (e.g. from KEGG) and protein interactions to uncover shortest path links between pathways altered by diet [165].

Epigenomic studies in these contexts have assayed CpG DNA methylation, chromatin accessibility, and individual transcription factor binding profiles. Li *et al.* (2013) compared CpG methylation patterns in the livers of mice born from lean or obese/diabetic mothers using arrays and found changes in methylation patterns near genes associated with development [166]. A study by Nilsson *et al.* (2015) collected liver biopsies from 35 diabetic and 60 non-diabetic humans and analyzed CpG methylation with HumanMethylation450 BeadChips [167]. They found 251 differential CpG sites, 94% of which were lower in methylation in diabetic patients, that mapped to some interesting gene candidates. Leung *et al.* (2014) profiled mouse hepatic epigenomes using FAIRE-Seq and histone modification ChIP-Seq following HFD and used these data to map changes in active chromatin to gene expression changes measured by RNA-Seq. They also inferred bound transcriptional regulators in these regions, finding enrichments for the liver factors HNF4α, C/EBPα, and FOXA1 [168]. Our group and others have directly assessed the genome-wide binding profiles for a number of such liver factors in mice and humans [169-173]

The application of quantitative, high-throughput proteomics in these contexts has been gaining prominence in recent years. Deng *et al.* (2010) measured diabetic rat liver mitochondrial total, phospho, and hydroxy proteomes using shotgun methods and found evidence for up-regulation of proteins involved in fatty acid β-oxidation, TCA cycle, and oxidative phosphorylation, along with depression of anti-apoptotic and anti-oxidative stress proteins [174]. Guo *et al.* (2013) also applied quantitative shotgun proteomics to mitochondria, but following HFD in mice, and found up-regulation by HFD of proteins involved in similar biological processes [175]. A study by Sabido *et al.* (2013) used targeted proteomics (by selected reaction monitoring) to quantify 144 proteins involved in insulin signaling and general metabolism in the livers of C57BL/6J and 129Sv mice following 6 and 12 weeks of HFD [176]. They observed robust early responses to diet and noted distinct proteomic profiles between the two mouse strains that separated by expression levels for metabolic enzymes involved in the TCA cycle, β-oxidation, fatty acid biosynthesis, and glycogen metabolism. Wu *et al.* (2014), also using selected reaction

monitoring, profiled 192 liver metabolic proteins across 40 strains of BXD mice fed chow and HFDs [177]. In addition, they profiled mRNA expression by microarray and found that changes in mRNA and protein expression due to diet are only modestly correlated (r ~0.31) and that 80% of all observed expression and protein quantitative trait loci (eQTL and pQTL) are distinct to either transcripts or proteins.

Metabolomics is also increasingly being applied in the type 2 diabetes research field [178]. Kim *et al.* (2011) measured liver and serum metabolites from lean and obese mice; partial least squares discriminant analysis (PLS-DA) revealed that fatty acids, lipid metabolism intermediates, amino acids, and monosaccharides separated the two conditions [179]. Wang *et al.* (2011) used targeted LC-MS/MS to measure metabolite levels in 2,422 individuals over a 12 year period, 201 of which developed diabetes, and found that elevated levels of branched chain amino acids are predictive of diabetes risk [180]. Additional studies have characterized 2-aminoadipic acid [181], glycine [182], and glyoxylate [183] (among others) as marker metabolites for type 2 diabetes. Considerable effort has been devoted as of late to studying the metabolomic profiles within plasma and some tissues of humans and mice in order to characterize features and risk factors for type 2 diabetes.

The majority of the omic studies discussed thus far focused on individual regulatory layers or did not focus on integrative analyses that jointly considered multiple datasets. The Wu *et al.* (2014) study described above indeed performed a targeted multi-omic analysis of hepatic transcription and protein expression, comparing changes induced by HFD at both regulatory levels [177]. Kirpich *et al.* (2011) also collected hepatic transcriptomic and proteomic data from mice following HFD but mostly described analyses of both data types in isolation [184]. Meierhofer *et al.* (2014) collected transcriptomic, proteomic, and metabolomic data from the white adipose tissue and livers of chow and HFD-fed mice [185]. They performed gene set enrichment analysis (GSEA) on individual datasets, combined enrichment analysis on all three (using the IMPaLa web tool [186]), and network analysis with protein interactions from the STRING database [187]. Their latter analysis was only performed on their proteomic data, which identified SDHB and SUCLG1 as important hub proteins in adipose tissue.

Studies attempting more explicit integrative assessment and modeling of multi-omic data have also been attempted to varying degrees in this context. Oberbach *et al.* performed independent component analysis on proteomic and metabolomic data obtained from normal and obese human plasma samples and found that sample discrimination was enhanced when the two data types were considered together [133]. Miraldi *et al.* used stochastic multivariate regression to predict altered lipid profiles from phospho-tyrosine proteomic data collected from the livers of HFD mice either expressing or lacking hepatic *Ptp1b* [135]. This analysis suggested roles for phosphoproteins involved in oxidation reduction in modulating polyunsaturated fatty acid and triglyceride metabolism. The CircadiOmics resource developed by Patel *et al.* (2012) maps normal and HFD-fed mouse metabolomic and transcriptomic data onto an interaction network built from known pathway, transcriptional regulatory, and protein-protein interaction data [188]. In a similar vein, growing numbers of tissue and cell-type specific genome-scale metabolic models, or GEMs, are being built to analyze omic information [189]. Tissue-specific gene expression information (from microarray or RNA-Seq) and proteomic data (from tissue immunohistochemistry data available in the human protein atlas [190]) are now being used to constrain such models. GEMs of human hepatocytes [191] and myocytes [192] have been built recently. Lee *et al.* (2016) expanded upon these methods by incorporating transcriptional regulatory networks and protein interaction information with GEMs to produce hepatocyte, adipocyte, and myocyte interaction networks [193]. They additionally built networks of lean and obese human hepatocytes and adipocytes from transcriptomic data and predicted obesity-induced alterations to metabolite concentrations, specifically identifying dysregulation of mannose metabolism that they subsequently validated by metabolic profiling of plasma.

This section reviewed prior published omic systems biology studies of obesity-induced insulin resistance and type 2 diabetes. The data and methods presented in this thesis expand upon this body of work.

1.8. Overview of thesis contents

The overall goal of my thesis work was to apply quantitative, multi-omic systems biology approaches to the study of obesity-induced hepatic insulin resistance. This entailed 1) the

collection of matched omic data from the livers of mice fed chow and high-fat diets across the transcriptome (mRNA and miRNA expression), epigenome (histone modifications, chromatin accessibility), proteome (total protein expression), and metabolome, 2) the analysis and interpretation of individual datasets, and 3) the integration of these multi-omic data through Bioinformatics and computational modeling approaches, particularly utilizing network modeling methods, and subsequent analysis of these results for further hypothesis generation and testing. This work was conducted as part of a collaborative, multi-disciplinary team of experimental and computational scientists at MIT (laboratories of Ernest Fraenkel, Forest White, and Douglas Lauffenburger), the University of Massachusetts Medical School (Roger Davis laboratory), and Harvard Medical School (Jarrod Marto laboratory). The work described in (1) was completed by the experimental arm of this collaboration. My specific contributions to this effort involved the completion of the work described in (2) and (3). The approaches described in this thesis surpass prior efforts in this realm in terms of the diversity of omic datasets collected and in terms of the level of simultaneous integration and modeling of such data.

Chapter 2 describes a comprehensive multi-omic systems biology analysis of obesity-induced hepatic insulin resistance. We collected transcriptomic, epigenomic, proteomic, and metabolomic data from the livers of chow diet (CD) and 16 week HFD-fed mice. I analyzed each of these datasets individually and uncovered changes induced by HFD. Additionally, I compared and contrasted information gleaned from each individual omic dataset against one another. I then adapted an existing computational modeling tool, namely the prize-collecting Steiner forest (PCSF), to simultaneously incorporate this molecular information into a tractable network model describing dysregulated pathways and biological processes induced by HFD. This effort required the novel incorporation of protein-metabolite interaction information with known protein-protein interactions. I also implemented strategies to enhance the selection of network components with high specificity to this biological problem by dissuading the inclusion of network "hubs" in models, and describe methods that aid model selection among families of related solutions. My modeling efforts uncovered both well and poorly characterized aspects of obesity-induced hepatic insulin resistance, and our group performed follow-up experiments on hepatic tissue obtained from additional CD and HFD mice to validate alterations to specific biological processes, including hepatic architecture, bile acid metabolism, and apoptosis.

Chapter 3 describes a study in which we collected transcriptomic data from the livers of CD, 6 week, and 16 week HFD mice treated without and with the type 2 diabetes drug metformin. I analyzed the effects of these diets on hepatic transcription and compared expression profiles between mice treated with and without metformin. Additionally, we stimulated CD and 16 week HFD mice with intraperitoneal insulin to profile transcriptional changes induced by this hormone in both diets. I found that CD mice showed a robust transcriptional response to insulin, whereas this response was generally blunted in HFD mice. However, we observed a set of more than 100 genes that were altered by insulin specifically in HFD livers. Among these were regulators of G-protein signaling (RGS) genes. We performed follow-up studies on a particular gene, *Rgs4*, to confirm this effect of insulin and to further characterized the role of this gene in HFD livers.

Chapter 4 describes a study of the hepatic transcriptomes and epigenomes of mice fed chow, 16 week high-fat, and calorie-restricted (CR) diets. I analyzed the transcriptional changes induced by HFD and CR versus CD and found that both diets induce extensive gene expression alterations. Interestingly, I found a significant sub-set of genes modulated by both HFD and CR that change in the same direction compared to CD. We also collected DNase-Seq data to profile chromatin accessibility in these livers and I used motif analysis to identify transcriptional regulators that are likely associated with genes modulated by diet. Based on these results, we chose to further test the roles of two specific transcriptional regulators, PPARα and RXRα, by performing ChIP-Seq experiments for these factors in HFD and CR livers. We found extensive binding of these regulators near genes modulated by diet and specifically highlighted binding near genes involved in glucose metabolism. We further tested the role of PPARα in liver by treating mouse primary hepatocytes with the PPARα activator fenofibrate and found that this factor modulates anaerobic glycolysis. We additionally validated novel predicted target genes of PPARα by measuring gene expression changes following *in vivo* fenofibrate treatment.

Chapter 5 examined mouse hepatic micro RNA (miRNA) expression changes induced by 6 and 16 week HFD and describes methods for integrated analysis of miRNA expression, mRNA expression, and epigenetics. I found that HFD progressively alters the expression landscape of miRNAs in the liver. I developed an enrichment scheme to prioritize miRNAs that considers

overrepresentation of target genes modulated by HFD that are also predicted targets of each differential miRNA. We also used a network modeling algorithm that incorporated miRNA, mRNA, and epigenetic data to specifically probe miRNA-transcription factor interactions. Both methods prioritized miRNAs with both known and potentially novel regulatory roles in the context of hepatic insulin resistance.

Appendix A contains a manuscript describing the OmicsIntegrator software package developed by our lab. This package consists of two tools: 1) "Forest," which runs the PCSF algorithm on omic data against an input interactome and 2) "Garnet," which infers important transcriptional regulators from epigenomic, motif, and gene expression data. Variations of these methods were utilized in Chapter 2.

CHAPTER 2

HEPATIC DYSFUNCTION CAUSED BY CONSUMPTION OF A HIGH-FAT DIET

Obesity is a major human health crisis that promotes the development of insulin resistance and, ultimately, type 2 diabetes. The molecular mechanisms that mediate this response occur across many complex levels of biological regulation that are poorly understood. Here we present a comprehensive study of the liver in mice fed a high-fat diet. We used an integrative network modeling approach to interrogate the hepatic epigenomes, transcriptomes, proteomes, and metabolomes altered by this diet. Our analysis highlights disruption of the hepatic architecture and hepatocyte apoptosis as processes that contribute to liver dysfunction and low-grade inflammation during the development of diet-induced metabolic syndrome.

This chapter presents work that is currently under review as a manuscript.

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$^{\Delta}$, and Ernest Fraenkel$^{\Delta}$

*Denotes equal contribution. $^{\Delta}$Denotes equal contribution.

Author contributions: Conceptualization and methodology, A.R.S., N.J.K., X.X., D.A.L., F.M.W., R.J.D., and E.F.; Software and formal analysis, A.R.S.; Investigation, N.J.K., X.X., F.Z., S.B.F., Y.S.Y., and B.J.M.; Resources, R.J.D., E.F., and J.A.M.; Writing, A.R.S., E.F., and R.J.D.; Visualization, A.R.S. and N.J.K.; Supervision, E.F., R.J.D., F.M.W., and D.A.L.; Project administration, R.J.D.

Acknowledgments: We acknowledge members of the MIT BioMicro Center for assistance with sequencing data collection. This work was supported by R24 DK-090963 (E.F., R.J.D., F.M.W., and D.A.L.), R01 NS-089076 (E.F.), R01 DK107220 (R.J.D.), P01 NS047572, P50 HG004233, Dana-Farber Strategic Research Initiative, Juvenile Diabetes Research Foundation (J.A.M.), and used computing resources funded by the National Science Foundation under Award No. DB1-0821391 and sequencing support from NIH (P30-ES002109).

Accession numbers: The mRNA-Seq and ChIP-Seq raw and processed data are deposited in the Gene Expression Omnibus (GEO) as accession number GSE77625.

2.1. INTRODUCTION

Human obesity is a major world-wide health crisis [5], promoting metabolic syndrome, which is characterized by insulin resistance, hyperglycemia, and hypertension [7], together with β-cell dysfunction and ultimately type 2 diabetes [8]. The liver is an insulin-sensitive organ that is critical for the maintenance of normal glucose homeostasis [41]. Insulin promotes increased uptake of glucose in peripheral tissues (primarily skeletal muscle) and reduces hepatic gluconeogenesis [15]. Insulin resistance suppresses these normal regulatory mechanisms and

thus promotes hyperglycemia. Consumption of a high-fat diet (HFD) causes insulin resistance, which prevents insulin-mediated inhibition of hepatic gluconeogenesis [30]. Moreover, peripheral insulin resistance (e.g. in adipose tissue) causes increased lipolysis that promotes hepatic gluconeogenesis [46-48]. The critical role of the liver in glycemic regulation is particularly highlighted by the widespread use of the drug metformin to treat type 2 diabetes, which principally acts in the liver to inhibit gluconeogenesis and reduce plasma triglyceride levels [24]. Thus, understanding the molecular mechanisms of hepatic insulin resistance may provide a basis for the design of therapeutic interventions.

The intracellular pathways that promote and maintain insulin resistance and type 2 diabetes are complex. For instance, genome-wide association studies (GWAS) have shown that genetic variations in the genes encoding the insulin receptor substrate IRS1 and the transcription factors TCF7L2 and PPARG (to name a few) are strongly associated with human risk for type 2 diabetes [64]. As a result, systems biology approaches are increasingly being recognized as vital to the study of metabolic diseases [63]. Systems biology embraces the inherent complexities of disease and draws upon the wealth of available knowledge from molecular biology and biochemistry to facilitate comprehensive, multi-dimensional analysis and modeling of disease-relevant systems and processes [65].

New omic technologies enable rapid and comprehensive analysis of many biological regulatory levels. Epigenomic and transcriptomic methodologies (e.g. ChIP-Seq, mRNA-Seq) rapidly profile full genomic regulatory and gene expression landscapes [80]. Proteomic analysis via mass spectrometry is increasingly becoming more sensitive and comprehensive, allowing for detailed analysis of global and modified proteomes [194]. Metabolomics, the collective study of small molecule species, is now being used extensively to identify new mechanisms and biomarkers of metabolic disease in both targeted and untargeted fashions [118].

A few studies have attempted to analyze multiple types of omic data in the context of metabolic disease. Some have used statistical routines, such as correlation network analyses [133] or stochastic multivariate regression [135]. Other methods have focused on known pathways, overlaying proteomic and metabolomic data onto genome-scale metabolic reconstructions [154]

or combining mouse transcriptomic and metabolomic data with known pathway and regulatory data to allow exploration of local interaction neighborhoods around genes or metabolites of interest [195]. Here, we go beyond these prior methods by developing a framework that integrates matched multi-omic data into a tractable network model. Our approach is not biased towards analysis of interactions that occur within well-established signaling or metabolic pathways alone. Instead, we collate diverse types of interactions from databases of literature-curated and high-throughput data to build a large network of physical associations. We then use advanced network optimization methods to prune the possible interaction space to only the most relevant connections that model the input data. Our results are thus more interpretable and provide clearer directions toward follow-up study.

We present a large-scale, integrative systems biology study of high-fat diet (HFD)-induced hepatic insulin resistance. We fed male C57BL/6J mice a normal chow diet (CD) or a 16 week HFD to induce obesity and insulin resistance. We then collected multiple omic datasets from the livers of these animals; specifically, we used histone modification ChIP-Seq to profile the epigenomes, mRNA-Seq to quantify the transcriptomes, and mass spectrometry to assess the global proteomes and metabolomes of CD and HFD livers. We identified genes, proteins, and metabolites altered between CD and HFD. By jointly analyzing the epigenomic and transcriptomic data, we predicted transcriptional regulators that likely influence gene expression changes between the diets. We then developed a network modeling approach based on the prize-collecting Steiner forest (PCSF) algorithm [158, 159] to analyze all the omic data in the context of known protein-protein and protein-metabolite interactions. For this purpose, we constructed a vast interactome of such associations and developed computational methods to avoid biases from well-studied, highly-connected proteins and metabolites. The PCSF model revealed a richly interconnected network of biological species and processes perturbed by HFD that could be divided into functional sub-networks. This analysis uncovered well-established features of hepatic insulin resistance, including glucose, lipid, and amino acid metabolism. Importantly, it also revealed novel and poorly characterized aspects of the condition, including hepatocellular injury, cell-cell interactions, extracellular matrix (ECM) organization, and apoptosis. Finally, we validated some of our global network modeling predictions with additional experiments on frozen liver sections from CD and HFD livers. We showed that HFD feeding leads to disrupted

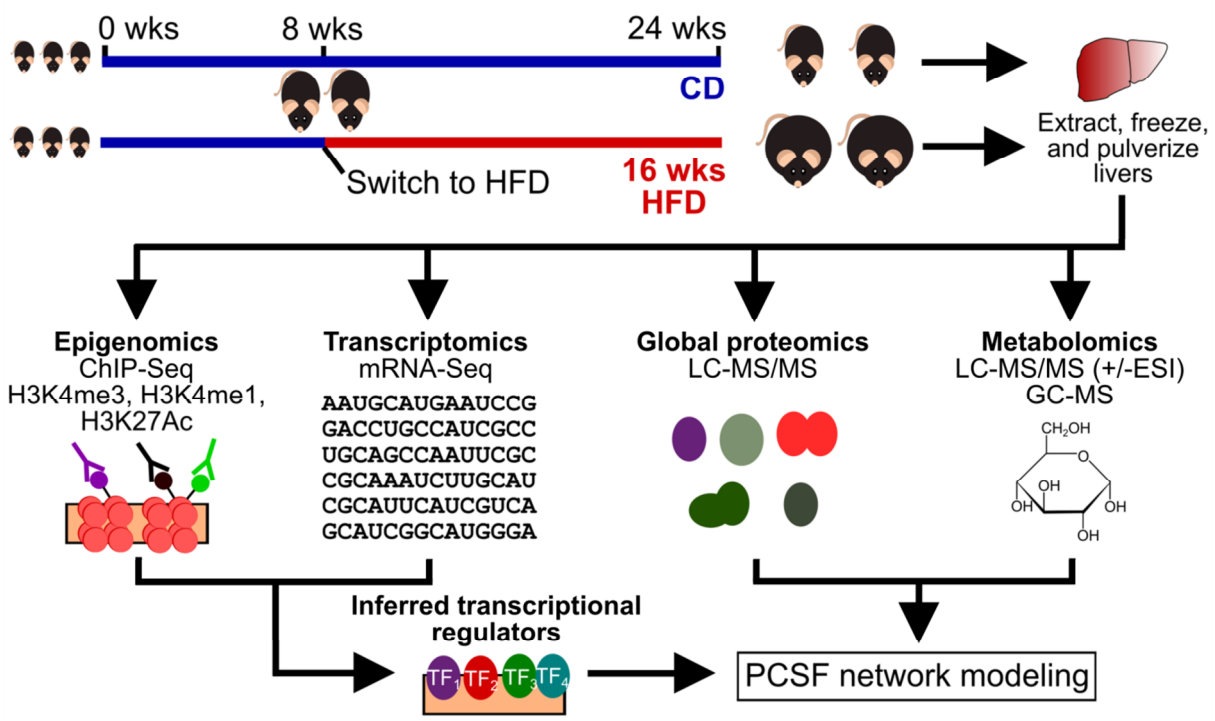


Figure 2-1. Overview of systems biology study of HFD-induced insulin resistance. We fed 8 week old male C57BL/6J mice a 16 week standard laboratory chow diet (CD) or a high-fat diet (HFD) to induce obesity and insulin resistance. At 24 weeks we sacrificed the mice and extracted, flash froze, and pulverized their livers. We used these tissue samples to assay epigenomes, transcriptomes, proteomes, and metabolomes. We then used mRNA-Seq (differential genes) and histone modification ChIP-Seq (valleys within enriched peaks) data with known DNA binding motifs to infer active transcriptional regulators. These regulators, along with differential proteins and metabolites, were used as input to the prize-collecting Steiner forest (PCSF) algorithm to uncover a network of interconnections amongst the data.

hepatic architecture and tight junctions, altered bile acid handling, and enhanced cellular apoptosis.

2.2. RESULTS

2.2.1. High-fat diet feeding induces obesity and insulin resistance in mouse

We examined diet-induced obesity and insulin resistance by feeding eight week old male C57BL/6J mice a HFD for 16 weeks (**Figure 2-1**). Control mice were fed a standard chow diet

(CD) for the same 16 week period and all animals were euthanized at the 24 week time point. This model is particularly suited for the study of human metabolic diseases as HFD consumption by mice induces complications consistent with the progression of human metabolic syndrome [62]. Indeed, we found that HFD-fed mice exhibited obesity, hepatic steatosis, hyperglycemia, insulin resistance, and glucose intolerance compared with CD-fed mice (**Figure 2-S1**).

2.2.2. Omic datasets demonstrate wide-ranging effects of HFD on mouse liver biology

We collected an array of datasets using high-throughput omic experimental methods to broadly capture the effects of HFD in the liver (**Figures 2-1 and 2-2**). We used the information obtained from analysis of these datasets to inform our subsequent integrative network modeling efforts.

Epigenomics: We profiled the epigenomes of CD and HFD livers with histone modification ChIP-Seq experiments for H3K27Ac, which marks active enhancers [91], H3K4me3, which marks active and poised promoters [90], and H3K4me1, which marks active and poised enhancers [91] (Figure 2-2, top panels). We tested for differences in histone modification levels between the diets but found few significant differential regions (< 1%). Overall, these data provide a comprehensive map of > 22,000 active regulatory regions in the liver genome.

Transcriptomics: We next collected transcriptomic data by mRNA-Seq to identify 2,507 genes differentially expressed between CD and HFD livers. Of these, 1,572 genes are up-regulated and 935 genes are down-regulated in HFD livers (**Figure 2-2, bottom left; Figure 2-S2A**). Genes up-regulated by HFD are enriched in lipid metabolism (*Aacs, Fasn*, *Ldlr,* and *Srebf1*) and carbohydrate metabolism (*Gck, Hk2,* and *Pfkl*) while genes down-regulated by HFD are enriched in amino acid catabolism (e.g. *Arg1*, *Gldc*, *Got1*, and *Hdc*) and small molecule catabolism (*Aadat, Aass, Cps1, Csad*). Shared biological enrichments between the two classes of genes include carboxylic acid and oxoacid metabolism. These genes and enrichment categories are generally consistent with prior data obtained from similar liver transcriptomic studies [56, 63]. We also performed TaqMan assays on additional CD and HFD samples (8 or more livers per condition) to further test for evidence of immune cell infiltration in HFD livers (as observed in our mRNA-Seq results) (**Figure 2-S3**). We found up-regulation of *Cd3e* (T cells), *Cd11c*

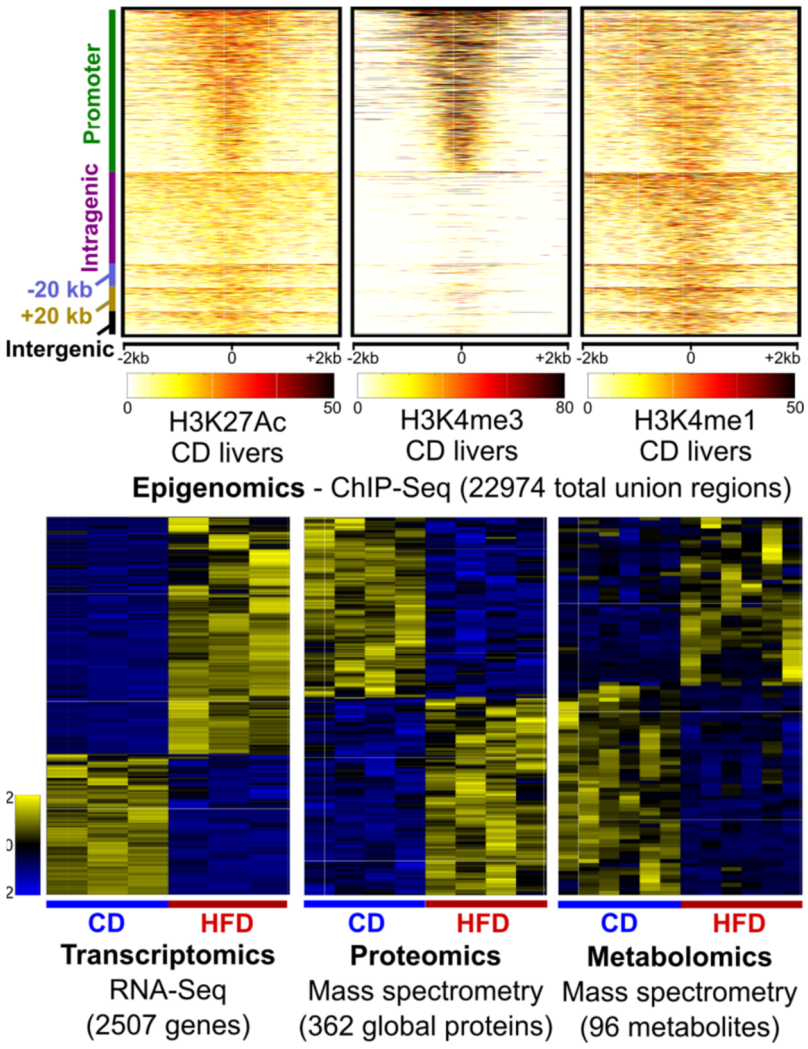


Figure 2-2. HFD induces perturbations to hepatic omic levels. (Top panels) Smoothed read density profiles in -/+ 2kb windows around the union of all identified enrichment regions (22,974 total) for histone marks H3K27Ac, H3K4me3, and H3K4me1 from CD liver samples. The mappings on left are with respect to the closest RefSeq gene start site: promoter (-2/+2 kb to start site), intragenic, -20 kb (within 20 kb upstream), +20 kb (within 20 kb downstream), and intergenic (>20 kb away from nearest gene). (Lower panels) We found 2,507 genes (n = 3 for CD and HFD), 362 global proteins (n = 4 for CD and HFD), and 96 metabolites (n = 6 for CD and HFD) perturbed by HFD consumption. Clustergrams show individual z-scored values for species from CD and HFD replicates. Only the most significantly changing peptide is shown as a representative for each of the differential global proteins, though full statistics were performed on all peptides.

(dendritic cells/monocytes/macrophages), *Emr1* (monocytes/macrophages), and *Nos2* (M2-like macrophages), together with down-regulation of *Arg1* (M2-like macrophages). These results

suggest immune cell infiltration indeed plays a role in promoting and maintaining the insulin resistant state of HFD mice.

Proteomics: We used mass spectrometry to quantify CD and HFD liver global proteomes, identifying 51,689 unique peptides that mapped to 6,384 unique proteins. We used a weighted least squares regression procedure to find 362 differentially expressed proteins, with 189 up-regulated and 173 down-regulated in HFD livers (**Figure 2-2, bottom middle; Figure 2-S2B**). Proteins up-regulated by HFD are uniquely enriched in fatty acid β-oxidation (e.g. CROT, ECI1, HADH), fatty acid transport (CD36, FABP1, FABP2) and carbohydrate biosynthesis (FBP1, GBE1, GCK, GYS2), while the proteins down-regulated by HFD are uniquely enriched in cholesterol biosynthesis (CYP51, DHCR7, FDPS, IDI1) and the urea cycle (CPS1, NAGS, OTC). Both sets of proteins are enriched in amino acid metabolism, carboxylic acid metabolism, and oxidation-reduction processes. Our findings are consistent with similar targeted proteome studies of HFD-induced changes in liver [177].

Metabolomics: We obtained metabolomic measurements by mass spectrometry of 381 metabolites in CD and HFD livers (**Figure 2-2, bottom right; Figure 2-S2C**). We found 96 metabolites that are significantly different between the two diets, with 43 up-regulated and 53 down-regulated by HFD. These metabolites include amino acids (11 up-regulated, 22 down-regulated by HFD), lipids (11 up, 21 down), carbohydrates (10 up, 1 down), and peptides (2 up, 2 down). We observed increased levels of glucose and other carbohydrate molecules; this was anticipated because hyperglycemia is a well-established feature of hepatic insulin resistance. The large number of gluconeogenic amino acids down-regulated by HFD are also consistent with reports from Zucker diabetic fatty rat livers [196].

The overall changes in gene and protein expression induced by HFD consumption are only weakly to moderately correlated (r = 0.2 – 0.4), even when we restrict our analyses to genes and proteins called significantly different between both conditions (**Figure 2-S4A-B**). The lack of correlation between protein and mRNA pairs in the absence of additional knowledge of translational and degradational rates has been observed in many other studies [106]. These findings are also consistent with results obtained from a smaller, targeted set of mRNAs and

proteins analyzed in CD and HFD livers (observed $r = 0.31$) [177]. We also observed specific biological processes that are enriched in the set of differential mRNAs but not in the differential proteins (and vice versa). For example, proteins up-regulated by HFD are uniquely enriched in fatty acid β-oxidation and carboxylic acid catabolism (**Figure 2-S4C**). These comparisons demonstrate how individual omic datasets can highlight different aspects of disease processes.

2.2.3. Epigenome and transcriptome dataset integration uncovers transcriptional regulators influencing differential gene expression

We collected epigenomic and transcriptomic data with the goal of uncovering changes in transcriptional regulation between CD and HFD livers. To reconstruct this transcriptional regulatory network, we inferred the genomic binding locations of potential transcriptional regulators using our ChIP-Seq datasets and DNA binding motif data from TRANSFAC® [129]. As we found little evidence for changes in these histone modifications between diets, we used the set of significant ChIP-Seq regions in CD livers for our analyses. We searched each dataset for histone "valleys", or regions between peaks of local modification enrichment where histones are depleted and where regulators likely bind (**Figure 2-3A**), and merged these into one set of 123,974 total genomic loci. We then scanned the genomic sequences underlying these regions for matches to a set of 1,588 DNA binding motifs that map to at least one human or mouse transcriptional regulator (**Figure 2-3B**). For each regulator (motif) and each differentially expressed gene, a transcription factor affinity (TFA) score was derived as a distance-weighted sum of individual motif enrichment scores in regions near the gene's annotated transcription start site. We then used linear regression of each motif's TFA scores against the expression levels of all the differentially expressed genes and took significant regression coefficients (FDR < 0.01) as evidence for active regulators (**Figure 2-3C-D**).

In total, we identified 358 significant DNA binding motifs that mapped to 272 unique transcriptional regulatory proteins. Among these significant regulatory proteins are known liver-enriched transcription factors, including hepatic nuclear factors 1α, 1β, and 4α, retinoid X receptors α and β, peroxisome proliferator-activated receptor α, and C/EBPα [197, 198]. We also found strong enrichment for nuclear factor I proteins (A, B, C, and X), SOX4, FOXO1, and the

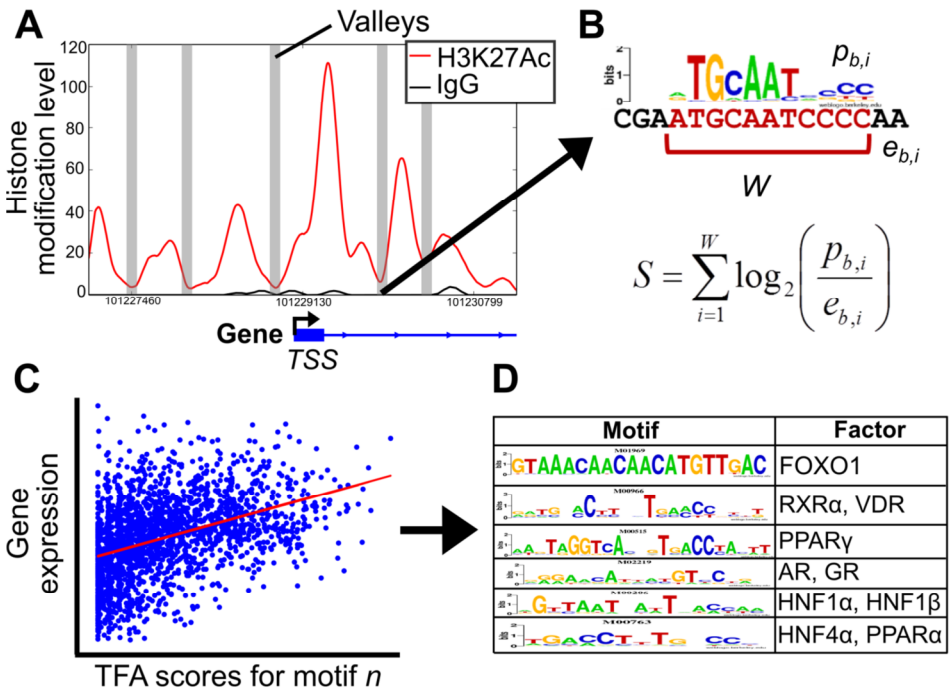


Figure 2-3. Motif regression procedure identifies transcriptional regulators. (A) We extracted read density profiles for significantly enriched histone modification levels, smoothed the profiles, and scanned for "histone valleys," or regions of local signal depletion (an H3K27Ac enrichment region is shown here as example). (B) For each valley, we scanned the underlying genomic sequence for matches to a library of DNA binding factor motifs. Against each differential gene, we computed a transcription factor affinity (TFA) score for all motifs as a distance-weighted sum of individual match scores. (C) For each motif, we used linear regression to predict gene expression levels from the motif TFA scores. (D) This procedure found 358 significant motifs that map to 272 regulatory proteins; select results are shown in the table.

vitamin D receptor (VDR). These significant factors served as the core transcriptional regulatory data that we incorporated into our network models.

2.2.4. Prize-collecting Steiner forest model integrates multiple omic data sets

Each type of omic data provides a glimpse into the effect of HFD on a particular regulatory level. To obtain a more comprehensive view of the data, we expanded upon an established network modeling algorithm called the prize-collecting Steiner forest (PCSF) [158, 159]. We built a combined protein-protein and protein-metabolite interactome from the iRefIndex (version 13) database [110] for protein-protein interactions and obtained protein-metabolite interactions from

the human metabolome database (HMDB, version 3.6) [117] and the human metabolic reconstruction Recon 2 (version 3) [76]. To account for the differences in reliability of the various types of interactions, we assigned to each an "edge cost" that scaled inversely with our confidence in the interaction (see Methods for details). We used this interaction network and the omic data as input to the PCSF algorithm to identify interactions that connect the omic data (**Figure 2-S5**).

As part of the PCSF approach, omic results (e.g. differential proteins) are assigned prizes (e.g. as $\log_2$ fold-changes) and the algorithm attempts to maximize the inclusion of these prize nodes while avoiding low-confidence edges, which have high edge costs. Thus, the algorithm is not constrained to include all data in the final network, but at the same time is capable of introducing species not present in the original set of data. These interactome-derived species, termed "Steiner" nodes, are included when necessary to fill connection gaps between the data. We also implemented a method that assigns "negative prizes" to interactome nodes with many interactions. These highly-connected species, referred to here as "hubs", have a high likelihood of appearing in network models run with almost any input data (e.g. ubiquitin, water). Negative prizes discourage the algorithm from using such nodes in the PCSF solution and allow for more specific interactions to explain the data (**Figure 2-S6A-B**).

We used as input data, or "terminals" in PCSF parlance, 83 differential metabolites, 329 differential proteins, and the 272 transcriptional regulators identified by our motif regression analysis. We sampled and merged multiple, related solutions to the PCSF problem by running the algorithm on the same data multiple times with small amounts of random noise added to the edge costs. This procedure produced a richer set of possible connections explaining the data and enabled assessment of individual network components' robustness. We also assessed how specific the nodes included in our final model are to hepatic insulin resistance by comparing how many times each node in the final solution appears in networks generated with random input data (i.e. nodes selected at random from the interactome that match the degree distribution of the real input data).

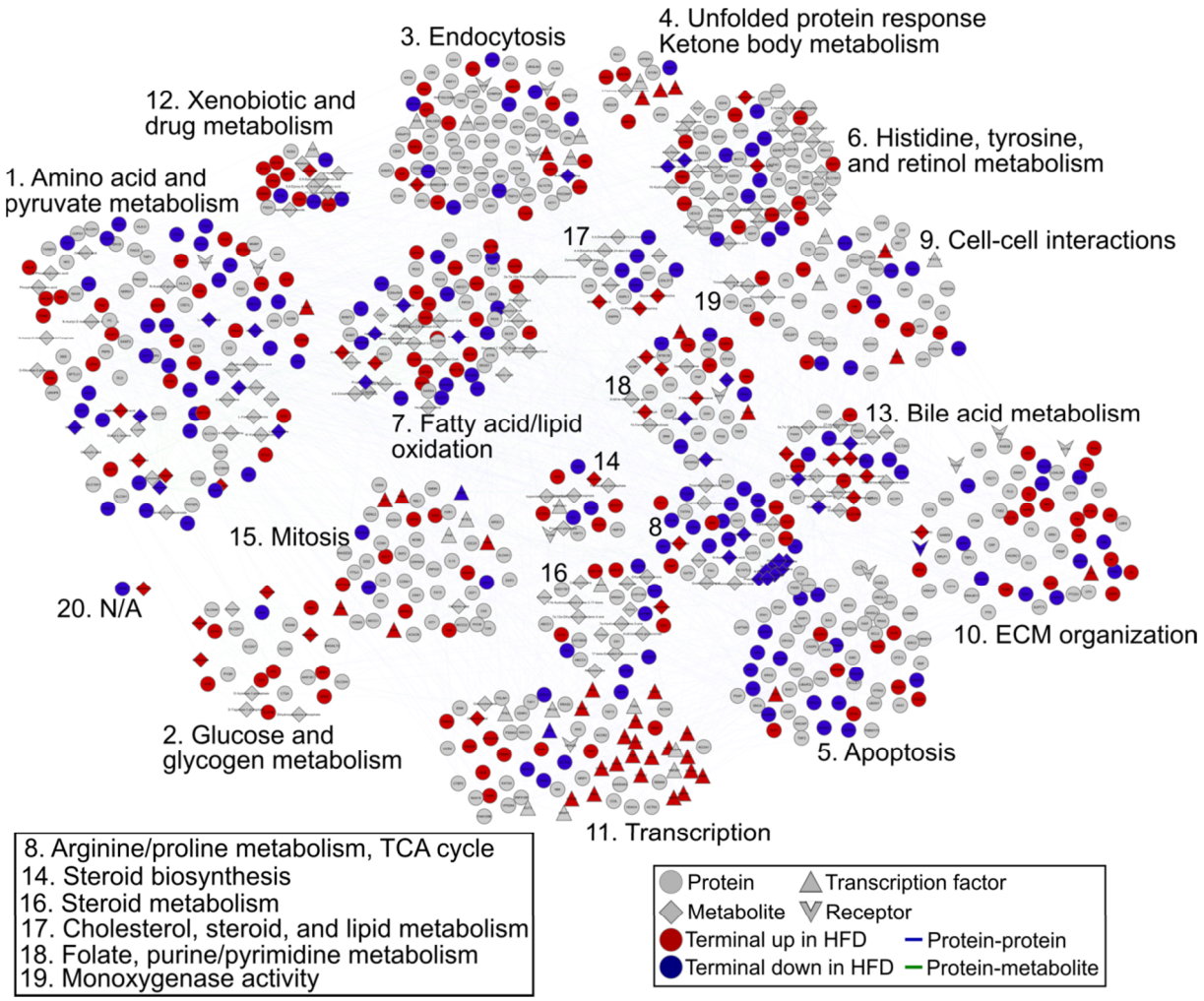


Figure 2-4. Multi-omic PCSF model uncovers features of hepatic insulin resistance. The full PCSF model includes 398 terminal nodes and 509 predicted Steiner nodes connected by 2,365 interactions. We divided the solution into 20 sub-networks and highlight the specific biological processes contained within these. Colored nodes (red or blue) represent terminal nodes, gray nodes represent Steiner nodes, and shapes indicate node types (proteins, metabolites, transcription factors, or receptors).

The full PCSF solution (**Figure 2-4**) includes 907 species connected by 2,365 interactions (also see **Table 2-1**). We found that the vast majority of nodes included in the final network are very specific to our particular problem (**Figure 2-S6C**). To increase interpretability of the network model, we identified smaller sub-networks and performed enrichment analyses on these using a variety of gene and small molecule ontology and pathway sets (**Figure 2-S7**). Additionally, we devised a scheme to rank interactome-derived Steiner nodes by their likely importance in the model according to several features, including the robustness and specificity of nodes. We used a

Terminal type	Number of terminals	Number included in final model	% included
Metabolites	83	63	75.9
Global proteins	329	301	91.5
Transcription factors	272	34	12.5

Table 2-1. PCSF model terminal node inclusion statistics

weighted summation of scores based on these features to perform this ranking (see Methods for details).

2.2.5. The PCSF model introduces species with known relevance to metabolic disease

We developed an automated strategy to determine which of the nodes in the network are expected in the context of the observed metabolic states and related diseases and which are potentially novel predictions. For this purpose, we used the DisGeNET database [199], which collates gene-disease information from public data as well as from literature via natural language processing tools, to determine which of the predicted molecules introduced by the PCSF into the network (Steiner nodes) are known to be associated with obesity, insulin resistance, and/or type 2 diabetes. Of the 394 protein Steiner nodes included in our model, 121 (~30%) possess some known disease link according to DisGeNET. Some examples include: clusterin (CLU), in which polymorphisms are associated with type 2 diabetes [200] and where knock-out in C57BL/6J mice exacerbates HFD-induced insulin resistance [201]; L-arginine:glycine amidinotransferase (GATM, aka AGAT), where knock-out in C57BL/6J mice depletes creatine, enhances glucose tolerance, and protects from diet-induced obesity (effects that, interestingly, can be reversed with oral creatine supplementation) [202]; and nuclear receptor co-activator 1 (NCOA1, aka SRC-1), depletion of which can result in increased glucose uptake, enhanced insulin sensitivity, and resistance to age-associated obesity and glucose intolerance [203]. Literature review revealed additional Steiner nodes with known relevance to disease, including the metabolite glyoxylic acid, which has been characterized as a marker metabolite for type 2 diabetes [183]. Thus, our model incorporates many predicted nodes with known relevance to these conditions, though there are still many whose roles are not well-established or have not yet been characterized in these contexts.

2.2.6. The PCSF model includes processes with known relevance to insulin resistance

We observed sub-networks enriched in glucose and glycogen metabolism (sub-network 2), amino acid metabolism (sub-network 1), fatty acid and lipid oxidation (sub-network 7), and transcriptional regulation (sub-network 11), all well-established aspects of hepatic insulin resistance (**Figure 2-S7**). In sub-network 2, up-regulated glucokinase (GCK) connects up-regulated D-glucose, D-fructose, and its own regulatory protein (GCKR). Several studies have demonstrated a role for altered GCK regulation and activity in glycemic dysregulation and diabetes [204-206]. Sub-network 1 includes many down-regulated amino acids (e.g. glycine and serine) and altered amino acid metabolism enzymes, including aminoadipate aminotransferase (AADAT) and aminoadipate-semialdehyde synthase (AASS). Several high-ranking Steiner nodes appear here, including glyoxylic acid and CNDP2 (or peptidase A). Sub-network 11 contains the majority of the transcription factors from our motif regression, including RXRα, PPARα, and VDR. A high-ranking predicted node in this sub-network is the thyroid hormone receptor (THRA), which is involved in potentiation of insulin signaling in *db/db* mice [207] and reduction of hepatic steatosis in *ob/ob* mice [208]. Additionally, the Steiner nodes NCOA6 and NCOR2 (aka SMRT) play roles in regulating insulin signaling and sensitivity [209, 210].

2.2.7. The PCSF model identifies biological features of obesity-induced hepatic insulin resistance

We found sub-networks enriched in biological processes not typically associated with hepatic insulin resistance. One such sub-network is enriched in extracellular matrix (ECM) organizational and structural proteins (sub-network 10, **Figure 2-5**). Proteins associated with the ECM in this sub-network include collagens 1A1, 1A2, and 6A1 (COL1A1/1A2/6A1), as well as endoglin (ENG), fibronectin 1 (FN1), intergrin α5 (ITGA5), and the TGF-β receptor 1 (TGFB1). At the center of this sub-network is FN1 which connects, among other nodes, most of the collagen proteins and ITGA5. Both ENG and TGFBR1 are predicted Steiner nodes connected through ITGA5. Several Steiner nodes in this sub-network rank very highly by our criteria, including CD79A, 5'-3' exoribonuclease 1 (XRN1), and clusterin (CLU).

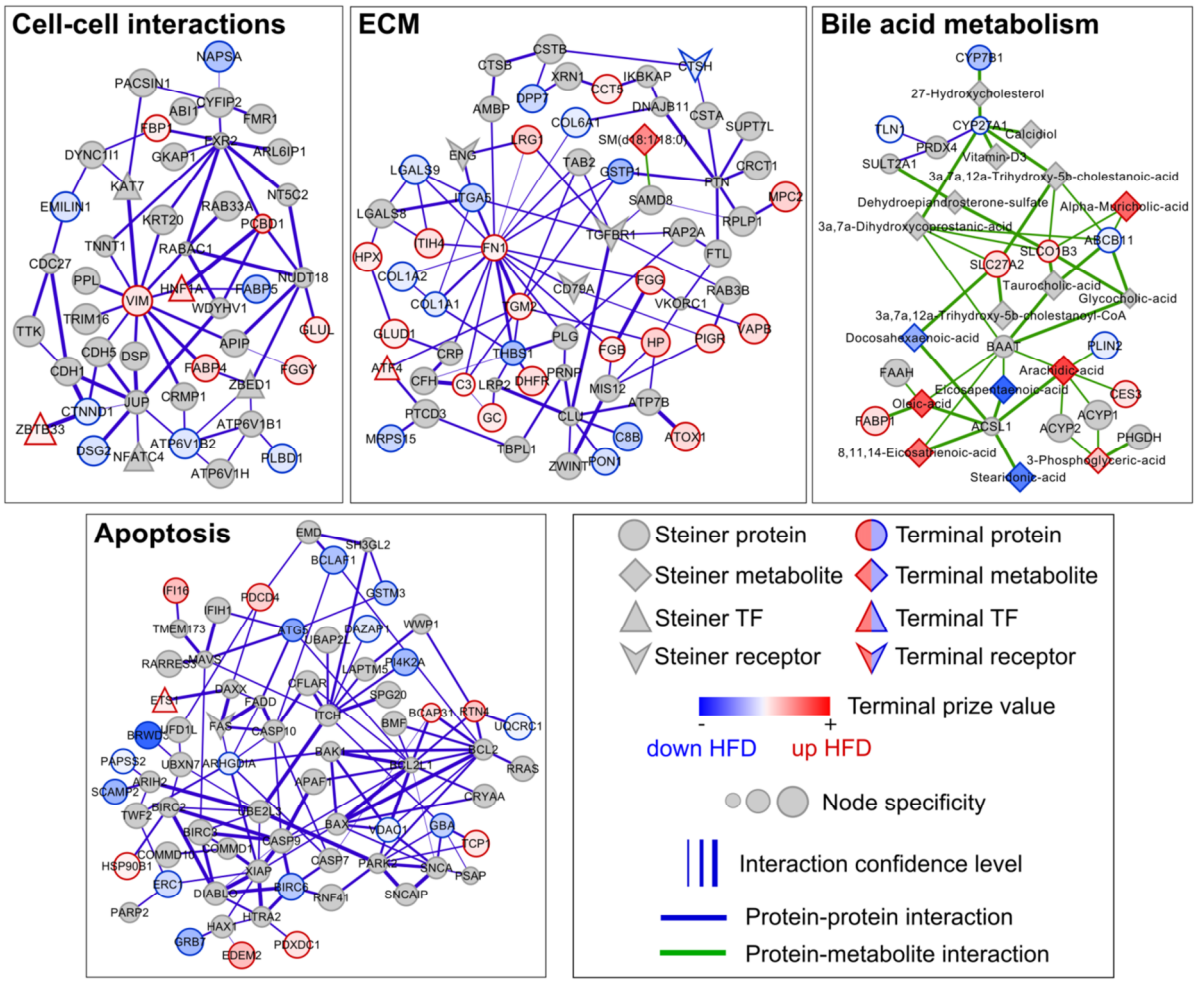


Figure 2-5. PCSF sub-networks for select biological processes. We highlight PCSF model sub-networks that are enriched in cell-cell interactions (top left), extracellular matrix (ECM, top middle), bile acid metabolism (top right), and apoptosis (bottom left). Note that node specificities should only be compared within sub-networks as overall panel sizes differ for clarity.

Changes to the hepatic ECM may also implicate altered cell-cell communication between hepatocytes in response to ECM and liver architectural disruption. Indeed, we found a sub-network enriched in proteins related to cell-cell interactions (sub-network 9, **Figure 2-5**). Included in this sub-network are the proteins E-cadherin (CDH1), cadherin 5 (CDH5), junction plakoglobin (JUP), and vimentin (VIM). These enrichments strongly suggest that changes to liver structure and the composition of the ECM are relevant to hepatic insulin resistance.

Another sub-network we identified is enriched in bile acid synthesis pathway members (sub-network 13, **Figure 2-5**), which include the terminals 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, activates 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 [211]. Bile acid coenzyme A (BAAT), a high-ranking Steiner node, conjugates these bile acids for biliary excretion [212], and is indeed connected to, among other metabolites, the Steiner nodes taurocholic acid and glycocholic acid. The terminal ABCB11 exports bile salts from hepatocytes [213] and SLCO1B3, a liver-specific organic anion influx transporter, transports bile salts, thyroid hormones, and eicosanoids [214]. ACSL1, an acyl-CoA synthetase that plays a role in lipid biosynthesis and fatty acid degradation, is also a high-ranking Steiner node by our scheme.

We also identified a sub-network enriched in apoptotic processes (sub-network 5, **Figure 2-5**). Terminal proteins involved in apoptosis here include autophagy related 5 (ATG5, a late apoptosis protein that interacts with FADD [215]), BCL-2-associated transcription factor 1 (BCLAF1), and IFN-γ-inducible protein 16 (IFI16). The majority of the apoptosis-related proteins are predicted nodes, including BCL2, BCL2L1, caspases 7, 9, and 10, FAS, the FAS-associated death domain (FADD), and BAD. The model captures aspects of the extrinsic apoptotic pathway, whereby the death inducing signaling complex composed of FAS, FADD, and pro-caspase 8 or 10 signals to downstream effectors [216], as well as the intrinsic pathway, which involves the pro-apoptotic Bcl-2 family member BAX and anti-apoptotic members BCL2 and BCL2L1 [217]. The model includes both initiator (CASP8, CASP10) and effector caspases (CASP7) linked to these initiator proteins [218]. Thus, our PCSF model overall suggests a role for apoptosis in maintaining hepatic insulin resistance.

2.2.8. Liver tissue analysis confirms global alterations in hepatic processes identified by the PCSF model

The network results imply roles for unexpected processes related to diet-induced insulin resistance. To test these predictions, we performed imaging studies on frozen liver sections from CD and HFD mice. First, we tested the prediction that HFD livers would display altered cell-cell interactions and overall structural deficiencies. We stained liver sections for Zo1, a cytoplasmic membrane protein of intercellular tight junctions, and cytokeratins 8 and 18, which are dimerized

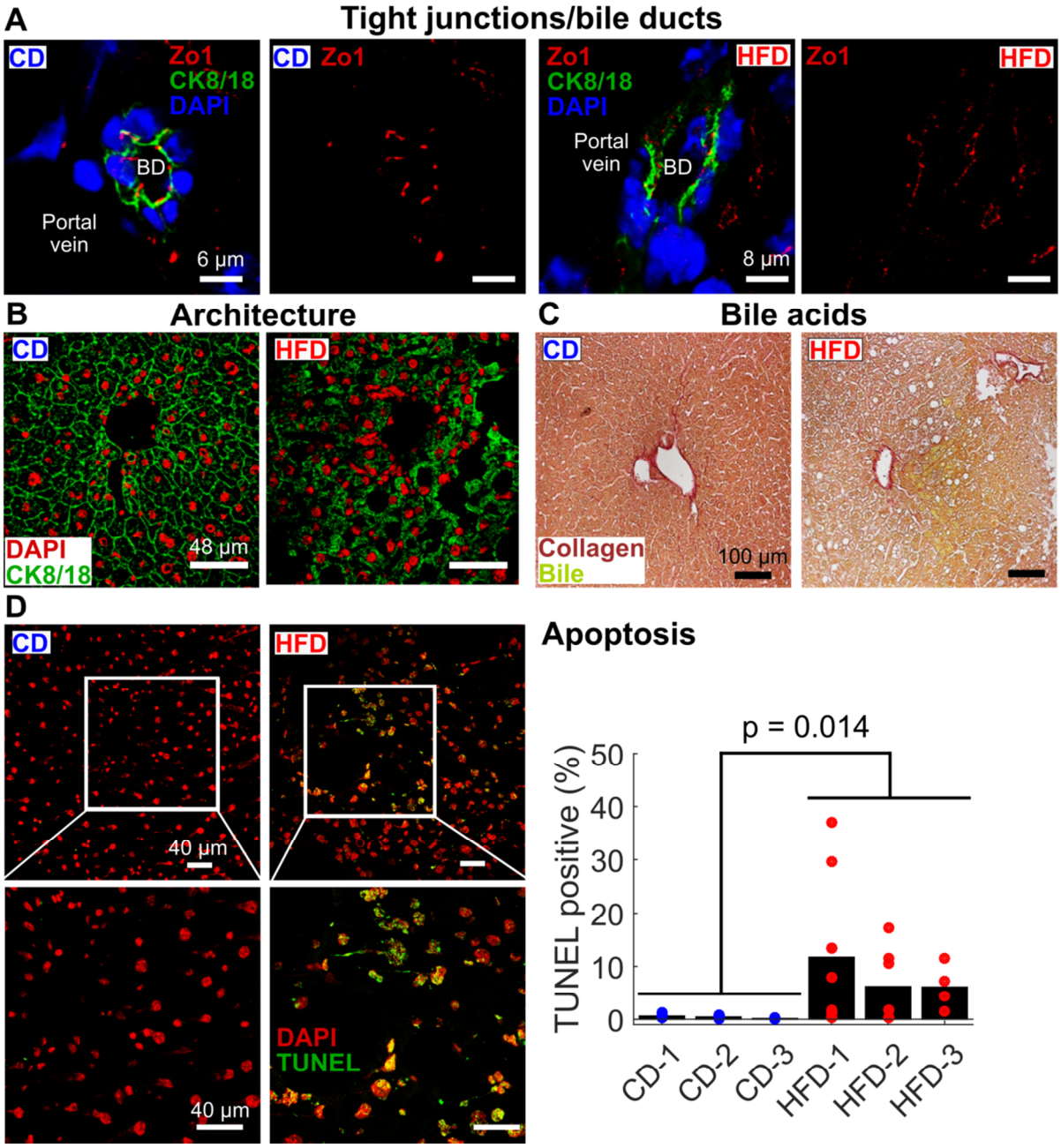


Figure 2-6. Hepatic imaging validates global PCSF model predictions.

Figure 2-6 (continued). Hepatic imaging validates global PCSF model predictions. (A) HFD induced changes in tight junction structure near bile ducts (BD) as assessed by cytokeratin 8/18 (CK8/18) and Zo1 staining. (B) CK8/18 staining revealed overall hepatic architectural defects in HFD samples. (C) We observed enhanced bile acid leakage in HFD livers stained for collagen and bile/bilirubin compared to CD. (D) TUNEL imaging revealed enhanced regions of hepatocyte apoptosis in HFD samples. Points on graph represent values from individual fields of view (n = 9, 7, and 4 for HFD livers; n = 4, 5, and 5 for CD livers) and bars are overall TUNEL positive fraction (total TUNEL positive cells over total cells) based on all fields of view. We found that the overall difference in TUNEL staining between the diets is statistically significant by two-tailed t-test (p = 0.014).

intermediate filaments present in epithelial cells that help maintain cellular structural integrity. Using DAPI staining to identify nuclei, we found cellular boundaries and tight junctions around bile ducts in the liver of CD-fed mice. By contrast, tight junctions and structure near bile ducts of HFD livers were highly disorganized (**Figure 2-6A**). In larger fields of view, we saw highly structured hepatocyte borders and normal architecture in CD livers (**Figure 2-6B**). In contrast, HFD livers displayed irregular cytokeratin 8/18 staining with few discernable cell borders, indicating overall disruption of the hepatic tissue architecture in response to the long-term dietary challenge.

We also tested the prediction that HFD livers would display abnormal bile acid handling by staining liver sections for collagen and bile/bilirubin (**Figure 2-6C**). As expected, we found no bile acid leakage or accumulation in CD livers. However, we observed significant bile accumulation in HFD livers. These results corroborate our prediction that HFD livers possess defects in bile acid maintenance and are consistent with the altered cellular structures we found surrounding bile ducts of HFD-fed mice.

Finally, we tested whether consumption of a HFD enhances the number of hepatocytes undergoing apoptosis in the liver. We used DAPI and terminal deoxynucleotidyl transferase dUTP nick end labeling (TUNEL) to assess the number of apoptotic cells. The fraction of TUNEL positive cells in CD livers was very low (~1%), whereas HFD livers displayed regions of high TUNEL positivity (as high as 37%, **Figure 2-6D**). While not prevalent in all regions of the livers, overall apoptosis was higher in HFD samples (**Figure 2-6D**, p-value = 0.014). Thus,

we show here evidence for enhanced hepatocyte apoptosis as a feature of HFD-induced hepatic insulin resistance.

2.3. DISCUSSION

Our large-scale integrative systems analysis of HFD-induced hepatic insulin resistance incorporated epigenomic, transcriptomic, proteomic, and metabolomic data. Using a network approach, we were able to highlight global biological processes perturbed by HFD. The algorithm also incorporated disease-relevant proteins and metabolites from the interactome that were either not measured or found to be differentially expressed in our omic data. We validated several high-level model predictions by examining livers for markers of specific physical features and biological processes. We found that HFD consumption perturbs hepatic architecture, disrupts bile acid handling, and enhances hepatocyte apoptosis.

The liver is a major contributor to overall glycemic regulation. Indeed, insulin-stimulated clearance of blood glucose is mediated, in part, by inhibition of hepatic gluconeogenesis [30], a fact highlighted by the widespread use of the drug metformin that targets the liver to lower blood glucose concentration in type 2 diabetic [24]. Consumption of a HFD causes hepatic insulin resistance, which prevents insulin-mediated inhibition of hepatic gluconeogenesis [30]. As expected, we found that the HFD feeding in mice caused obesity, insulin resistance, and impaired glucose homeostasis. The HFD also caused changes in >2,000 genes, 362 global proteins, and 96 metabolites.

Epigenomic data played an important role in identifying transcriptional regulators relevant to insulin resistance. We used a motif regression procedure with these data, mRNA-Seq data, and motif data to find likely transcriptional regulators. The top motifs that emerged from our approach are consistent with those identified in a prior study that used different epigenomic techniques [168], and both our study and theirs did not observe many changes in histone modification levels between the diets despite significant gene expression changes. An advantage of our integrative modeling approach is that even if a pathway is not detected as changing by one

experimental method such as ChIP-Seq, it may emerge in the network based on evidence from other types of data.

To integrate all the omic datasets, we built on the PCSF network modeling approach [158, 159]. The PCSF method is not required to include all omic data yet is capable of introducing critical predicted nodes important for establishing connections between the detected molecules. PCSF networks are generally much smaller and more tractable than solutions from more naïve methods and reveal interpretable sub-networks enriched in specific biological processes and pathways. Here, we have significantly expanded the scope of the PCSF methods by adding physical associations of proteins and metabolites to the protein-protein interactome. This unified approach allowed us to capture a wider range of biological pathways and processes relevant to insulin resistance. We employed several strategies to improve the accuracy of our networks, including penalizing highly connected ("hub") nodes, testing the networks for robustness to noise, and assessing the specificity of nodes to our particular data and problem.

Our integrated approach can identify many different types of links among the omic data. We found pathways that were largely dominated by proteomic data (e.g. cell-cell interactions, ECM, apoptosis), but also found several sub-networks almost entirely composed of protein-metabolite connections (e.g. bile acid metabolism, glucose metabolism). The inclusion of direct metabolomic data along with protein-metabolite interactions was critical to capturing, for instance, relevant connections among differential proteins whose roles are best explained in the context of metabolic processes (e.g. GCK, CYP7B1).

Increasingly, systems biology and omic approaches are being recognized for their utility to the study of insulin resistance and type 2 diabetes [63]. To date, however, few studies have formally integrated multiple types of omic data in these contexts, with even fewer including metabolomics. Prior studies attempting such joint analyses used correlative statistical routines [133, 135] or methods that overlay proteomic and metabolomic data onto genome-scale metabolic reconstructions [154]. The CircadiOmics resource maps metabolomic and transcriptomic data onto interactions derived from known pathway and transcriptional regulatory data, but lacks methods for identifying high-confidence sub-networks [195]. Our approach goes

well beyond these previous methods by incorporating multiple data types from the same samples, allowing for interactions that occur outside well-established signaling or metabolic pathways, and using advanced approaches to reduce the possible interaction space to only the most relevant connections, thus increasing the interpretability of results and providing clear guidance for designing experiments.

Our model uncovered a highly interconnected network associated with the insulin resistant state in the liver. We predicted that changes to the ECM, cell-cell interactions, and overall hepatic architecture are features of insulin resistance. Subsequent experiments confirmed that the overall structure of HFD mouse livers is highly disrupted, especially near bile ducts. Consistent with this observation, we also found enhanced bile acid leakage (cholestasis) into the tissue of HFD-fed mouse livers. These structural abnormalities likely also contribute to the increased apoptosis we observed in insulin resistant livers. The link between hepatic ECM and architectural structural remodeling with insulin resistance has been studied [219]. Indeed, tail vein injection of HFD-fed mice with a hydrolase for hyaluronan, an ECM component, reduces features of muscle and liver insulin resistance [220]. Moreover, integrin α1 subunit-deficient mice (*Itga1*$^{-/-}$) fed a HFD display reduced fatty liver content, but also severe hepatic insulin resistance, compared to wild-type HFD-fed controls [221].

The hepatic structural changes detected in HFD-fed mice may be related to changes in apoptosis. Crosstalk between proteins relevant to insulin resistance and hepatocellular injury, including TNF, NF-κB, and JNK, have been proposed as potential drivers of apoptosis in the liver [222]. Indeed, apoptosis is associated with severe hepatocellular injury and steatohepatitis [223]. Here we report increased hepatic apoptosis in HFD-fed mice. This increased hepatic apoptosis may be related to dysregulation of the hepatobiliary system [218] and promotes low-grade inflammation and hepatic insulin resistance.

To summarize, we undertook a large-scale systems biology approach to study HFD-induced hepatic insulin resistance. We integrated multiple types of omic datasets into a network model that uncovered altered biological processes associated with the condition. By incorporating metabolites into the protein-protein interaction network, we were able to identify a wide range of

molecular changes. We validated several global predictions from our network model with additional experiments and highlighted components relevant to the hepatic response to HFD consumption. The pathways and processes we found to be altered by HFD present a wide range of new directions for future research. Our methods are easily applicable to other large-scale omic analyses of diverse biological systems and diseases.

2.4. MATERIALS AND METHODS

2.4.1. Animals

We obtained male C57BL/6J mice (stock number 000664) from the Jackson Laboratories. All mice were housed in a specific pathogen-free facility accredited by the American Association for Laboratory Animal Care. We fed the mice a standard chow diet (Prolab Isopro RMH 3000, Purina) for 24 weeks or a high-fat diet (S3282, Bioserve) starting at 8 weeks for 16 weeks HFD. We measured fat and lean mass noninvasively using ^{1}H-MRS (Echo Medical Systems). We euthanized all mice at 24 weeks after an overnight fast and froze the livers prior to removal using clamps cooled in liquid nitrogen. The frozen livers were then pulverized into a powder using a CryoPREP impactor (Covaris). We prepared aliquots of pulverized liver for all samples for subsequent analyses. All experiments were carried out in accordance with guidelines for the use of laboratory animals and were approved by the Institutional Animal Care and Use Committees (IACUC) of the University of Massachusetts Medical School.

2.4.2. Glucose and insulin tolerance tests

We performed glucose and insulin tolerance tests by intraperitoneal injection of mice with glucose (1 g/kg) or insulin (1.5 U/kg) using methods described previously [224].

2.4.3. Immunoblot analysis

Protein extracts from pulverized liver were prepared in Triton lysis buffer (20 mM Tris [pH 7.4], 1% Triton X-100, 10% glycerol, 137 mM NaCl, 2 mM EDTA, 25 mM β-glycerophosphate, 1

mM sodium orthovanadate, 1 mM phenylmethyl-sulfonyl fluoride, and 10 μg/ml each of aprotinin and leupeptin). We quantified protein content by the Bradford method (Bio-Rad). Standard techniques were used to separate cell extracts (15-80 μg of protein) by SDS-PAGE for immunoblot analysis using antibodies from Cell Signaling (AKT and $pSer^{473}$-AKT). The primary antibodies were detected by incubation with anti-mouse or anti-rabbit IgG conjugated to infrared dyes (IRDye®, LI-COR Biosciences). We detected immune complexes using the Odyssey infrared imaging system (LI-COR Biosciences).

2.4.4. mRNA-Seq and 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 [122] (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) [225]. 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 2-S2A**). We used the Bioconductor package conditional quantile normalization (CQN, version 1.6.0) [226] to remove systematic biases due to GC-content and gene length coverage and used DESeq2 (version 1.0.18) [227] 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).

2.4.5. ChIP-Seq and analysis

Histone modification ChIP experiments were performed using the MAGnify Chromatin Immunoprecipitation System kit (Life Technologies, Carlsbad, CA) with antibodies against

H3K4me1 (17-676, Millipore), H3K4me3 (17-614, Millipore), and H3K27ac (ab4729, Abcam, Cambridge, MA). ChIP-Seq libraries were constructed using the NEBNext DNA Library Prep Master Mix Set for Illumina (New England Biolabs, Ipswich, MA) and sequenced on an Illumina Hi-Seq 2000 machine. We aligned raw reads using Bowtie (version 0.12.7) [228] and performed peak calling using MACS (version 1.4.0rc2) [229] against an IgG control. We considered significant MACS peaks to be those possessing a p-value < 1e-6 and an FDR < 10%. We also performed differential peak analyses between conditions of the same histone mark. We used MACS-called peaks on replicate-pooled samples and merged significant peak regions from each condition into one set of common genomic loci. We then extracted raw read counts in these regions from each individual replicate ChIP-Seq sample and used DESeq2 [227] to perform the differential enrichment analyses on the read counts. We considered regions possessing an FDR-corrected p-value < 0.05 as significant.

2.4.6. Global proteomics

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 [230]. 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 [231], 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 [232] 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 [233]). 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 [234] 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 2-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.

2.4.7. Metabolomics and analysis

We extracted and split samples (6 independent livers per condition, per Metabolon Inc. recommendations for appropriate statistical power) into equal parts for analysis on GC-MS and LC-MS/MS (+/- ESI) platforms (Metabolon Inc.). A total of 381 metabolites were identified and quantitated. We imputed missing values with a k-nearest neighbors procedure (k=10), normalized samples according to the procedure in [234], and tested for differences using two-tailed t-tests, correcting p-values for multiple hypotheses. We observed strong intra-sample correlations between CD (Pearson's $r > 0.923$) and HFD ($r > 0.85$) replicate abundances (**Figure 2-S2C**). Metabolites possessing and FDR < 0.1 were deemed significant.

2.4.8. Motif regression analysis

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 [229] 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® [129], represented as position-specific scoring matrices (PSSMs). We extracted the underlying genomic sequences from the histone valley regions and used TAMO [235] 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). 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.

2.4.9. Prize-collecting Steiner forest (PCSF) modeling

PCSF formulation: The prize-collecting Steiner forest [158, 159] 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 [110] as our source for protein-protein interactions, which consolidates information from a variety of source databases. We used the MIscore system [236] 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) [117] and supplemented these with additional manually curated interactions from the human metabolic reconstruction Recon 2 (version 3) [76]. 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 2-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 $\beta = 10$ for global proteins and metabolites, $\beta = 1$ for transcription factors, $\omega = 3$, $D = 7$, $\mu = 5e\text{-}5$ for global proteins and transcription factors, and $\mu = 1e\text{-}3$ 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 2-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 [237] 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) [238], and canonical pathway node sets provided by MSigDB [239] derived from various database sources. We visualized all networks with Cytoscape [240].

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.

2.4.10. Liver tissue section analysis and imaging

Histology was performed using liver fixed in 10% formalin for 24 hours, dehydrated, and embedded in paraffin. Dewaxed and rehydrated sections (7 µm) were cut and stained for bile acids (product # KTHBI, American Master Tech Scientific) or with hematoxylin & eosin (American Master Tech Scientific). Sections (7 µm) prepared from liver frozen in O.C.T. compound (Tissue-Tek) were stained with Oil-red-O (Sigma) to visualize lipid droplets. We acquired images using a Zeiss Axiovert 200M microscope. Liver architecture was assessed using frozen sections fixed with 4% paraformaldehyde and stained with an antibody to cytokeratin 8 (TROMA-1-c, DSHB, University of Iowa). Immune complexes were detected using anti-rat Ig conjugated to Alexa Fluor 488. Liver damage was assessed in frozen sections (7 µm) fixed with cold ethanol/acetic acid (2:1) using an *in situ* cell death kit (Roche). Bile duct architecture was assessed in frozen sections (7 µm) fixed with cold methanol by staining with antibodies to Zo-1 (sc-10804, Santa Cruz) and Cytokeratin 8/18 (sc-52325, Santa Cruz). Immune complexes were detected using anti-mouse Ig conjugated to Alexa Fluor 488 and anti-rabbit Ig conjugated to Alexa Fluor 633 (Life Technologies). DNA was detected by staining with DAPI (Life Technologies). Fluorescence was visualized using a Leica TCS SP2 confocal microscope equipped with a 405-nm diode laser.

2.4.11. TUNEL imaging analysis

We used CellProfiler (version 2.1.1) [241] 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.

2.4.12. Clustering and enrichment analyses

All hierarchical clustering analysis was done with the *clustergram* function in Matlab with Euclidean distance and average linkage. For enrichment analyses, we used custom Matlab code implementing the hypergeometric distribution for enrichment p-value calculations and used the Benjamini-Hochberg FDR procedure to correct for multiple hypotheses. In general, an FDR < 0.1 was deemed significant.

2.5. SUPPLEMENTARY FIGURES

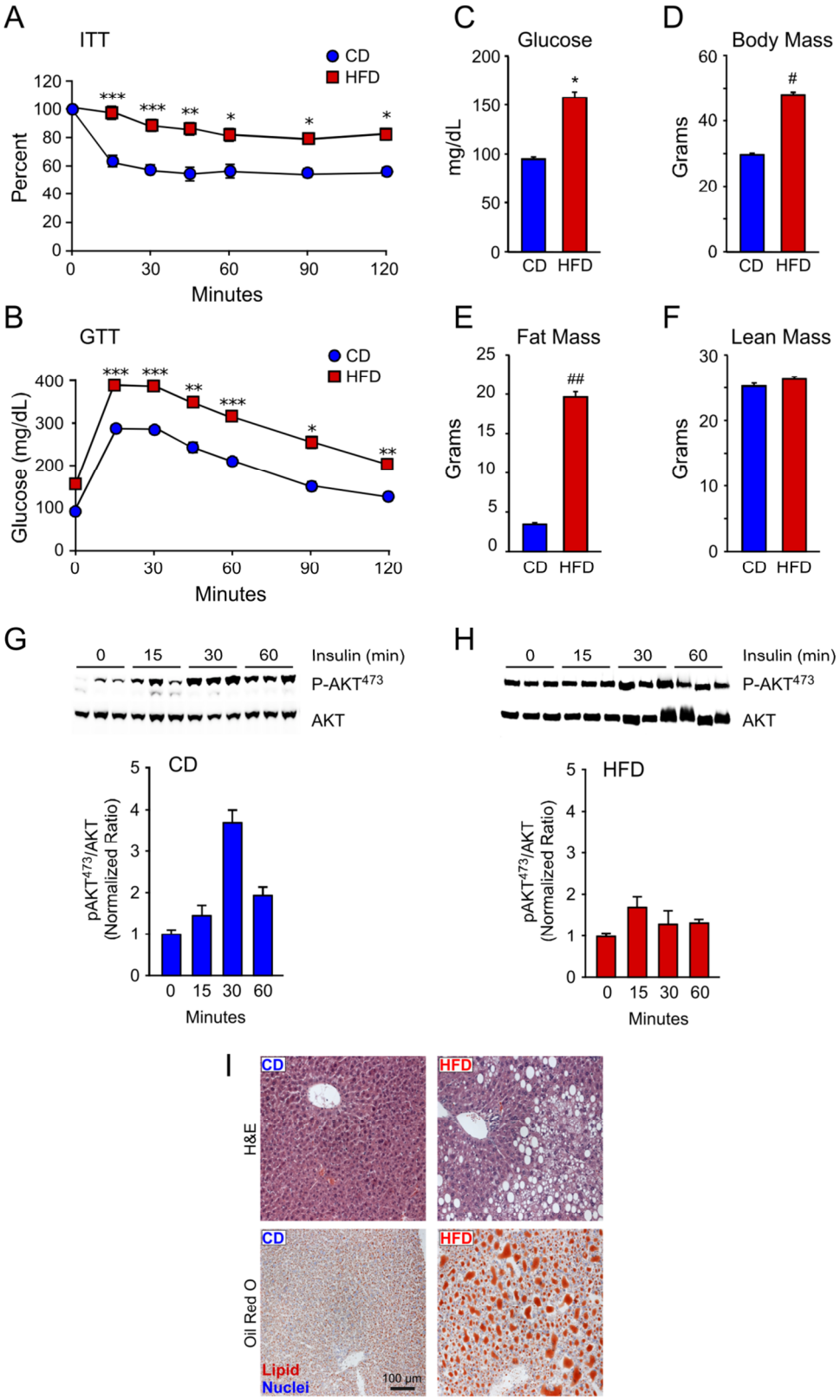


Figure 2-S1. Physiological analysis of CD-fed and HFD-fed mice and analysis of hepatic steatosis.

Figure 2-S1 (continued). Physiological analysis of CD-fed and HFD-fed mice and analysis of hepatic steatosis. (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 ^{1}H-MRS analysis (mean ± SEM; n=20; ##, $p < 5e\text{-}16$). (G, H) CD-fed (G) or HFD-fed (H) mice were treated with insulin (1 U/kg) by intraperitoneal injection. Hepatic AKT was examined by immunoblot analysis by probing with antibodies to $pSer^{473}$-AKT and AKT (mean ± SEM; n = 3). (I) 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 Oli Red O. The images are representative of sections prepared from three mice.

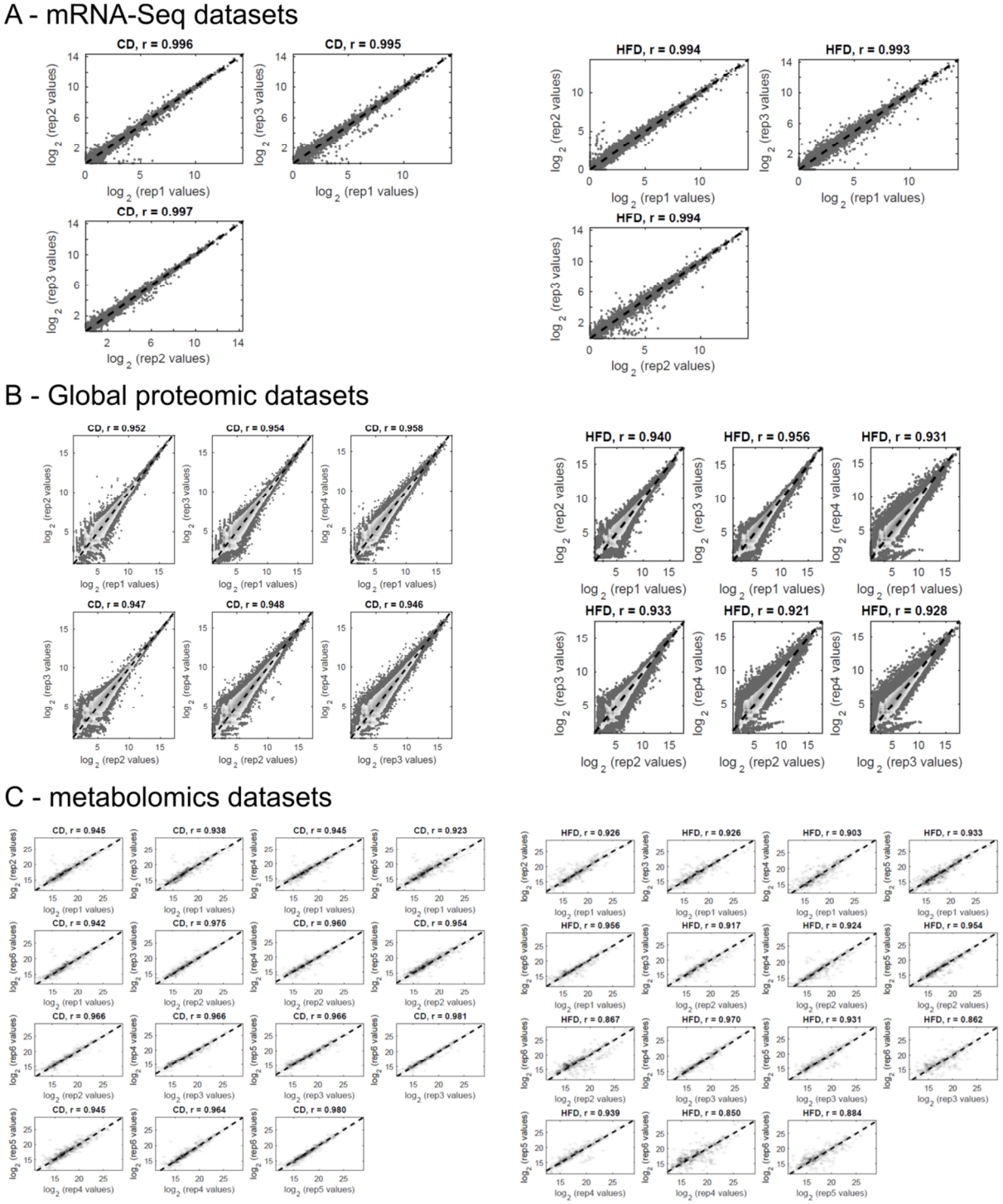


Figure 2-S2. Replicate correlations for mRNA-Seq, global proteomic, and metabolomic datasets.

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

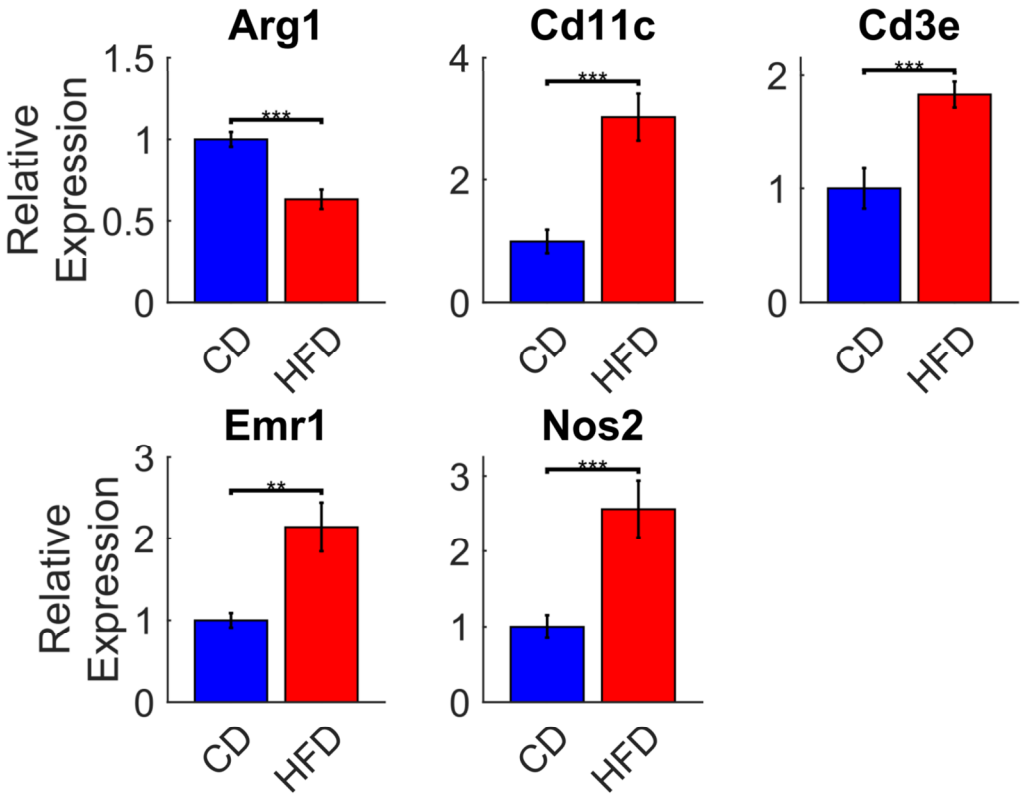


Figure 2-S3. TaqMan assays for immune cell marker genes. We performed TaqMan assays on immune cell marker genes to further establish evidence for the role of inflammatory processes in promoting and maintaining the insulin resistant state following HFD. All comparisons are between nine independent liver samples for each condition, except for *Emr1* which used eight livers per condition. *** $p < 0.001$, ** $p < 0.01$.

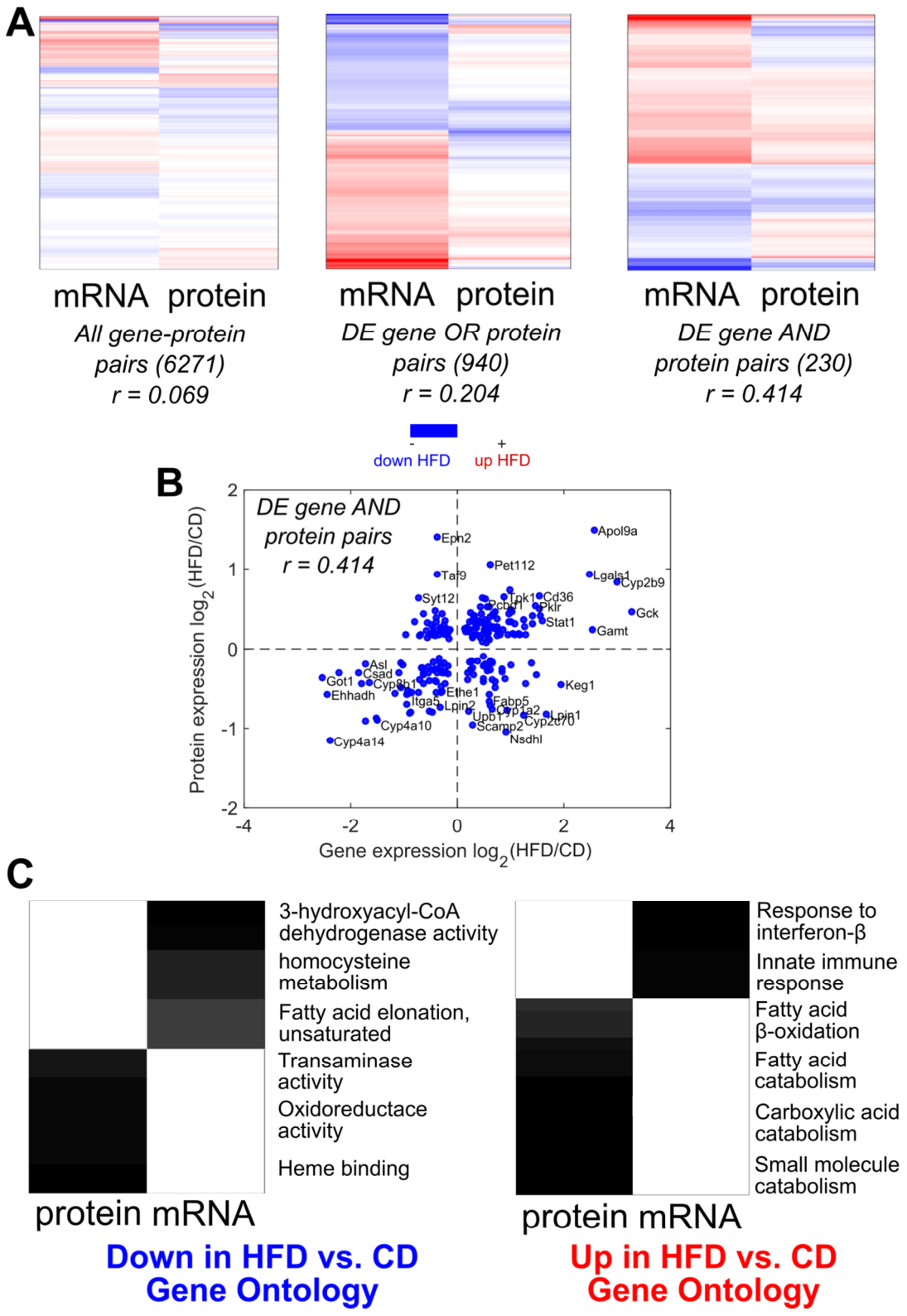


Figure 2-S4. Comparison of gene and protein expression changes between CD and HFD livers.

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

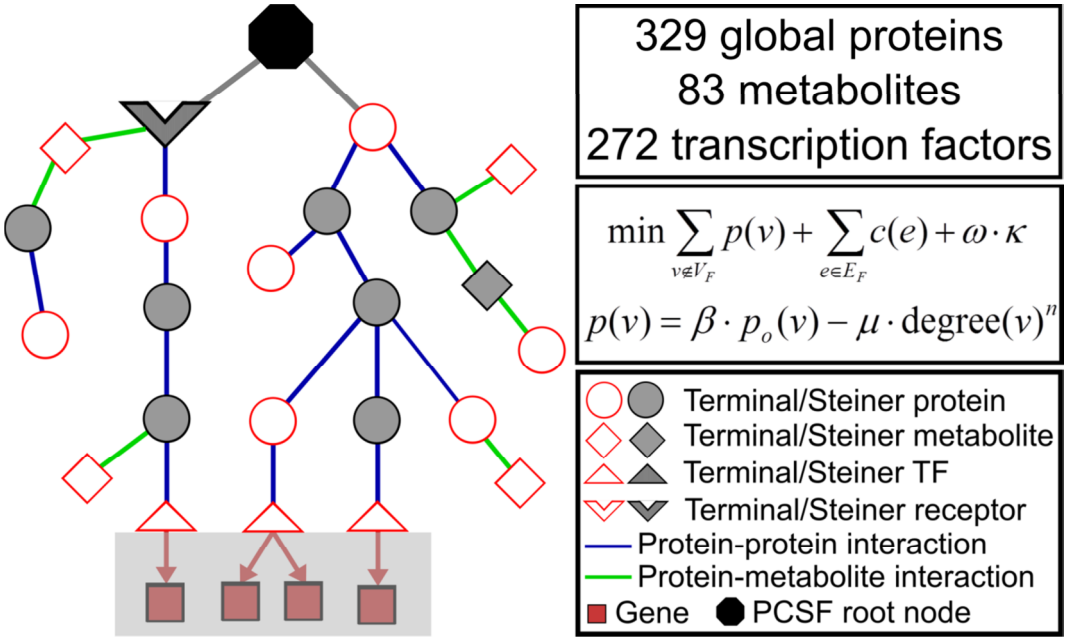


Figure 2-S5. Schematic overview of PCSF algorithm. The PCSF algorithm is initialized by connecting all terminal species (data in box at top right) via an artificial root node. A message-passing algorithm is then run to generate a network model that minimizes the overall objective function (middle box 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.

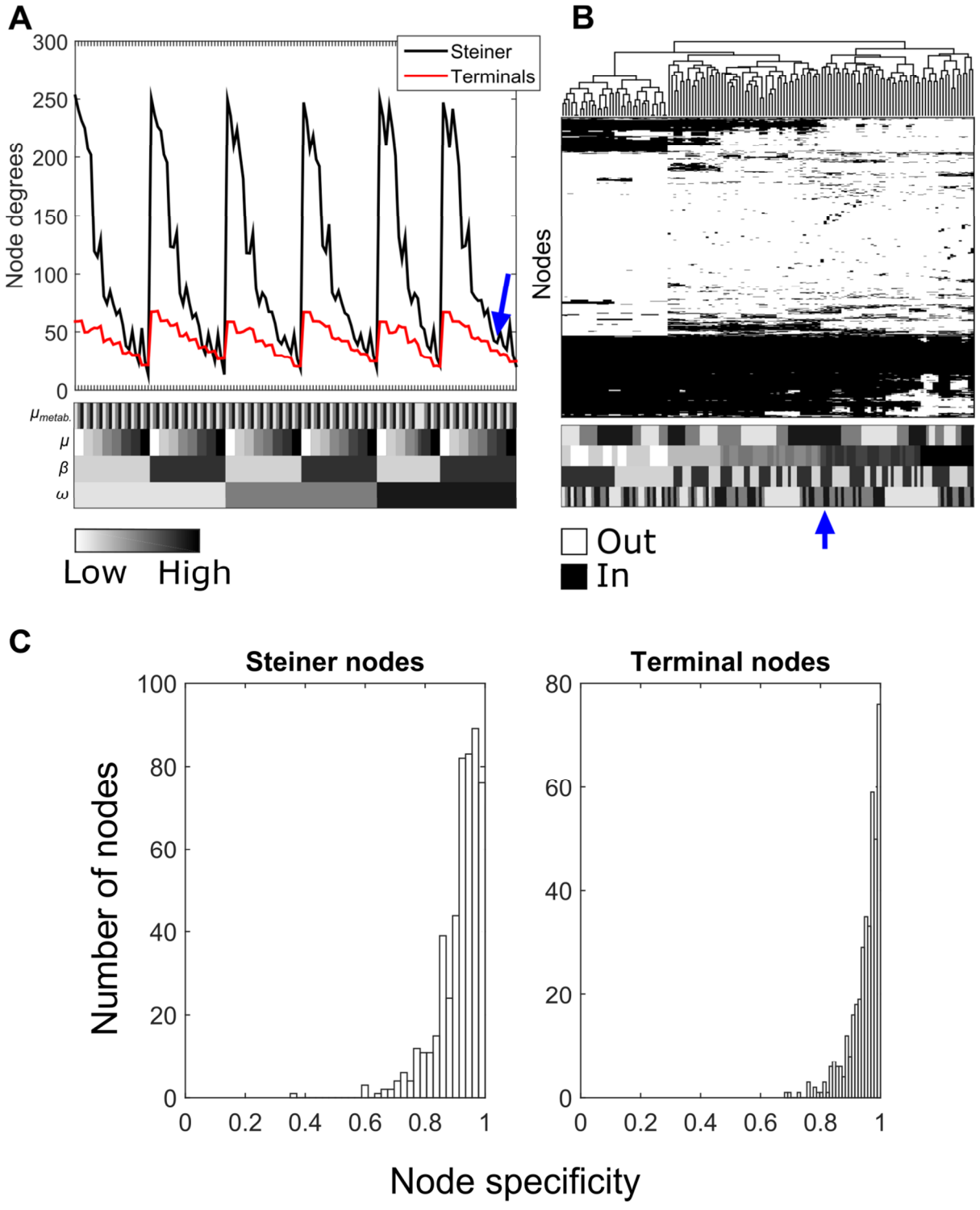


Figure 2-S6. PCSF model parameter selection criteria and final node specificities.

Figure 2-S6 (continued). PCSF model parameter selection criteria and final node specificities. 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.

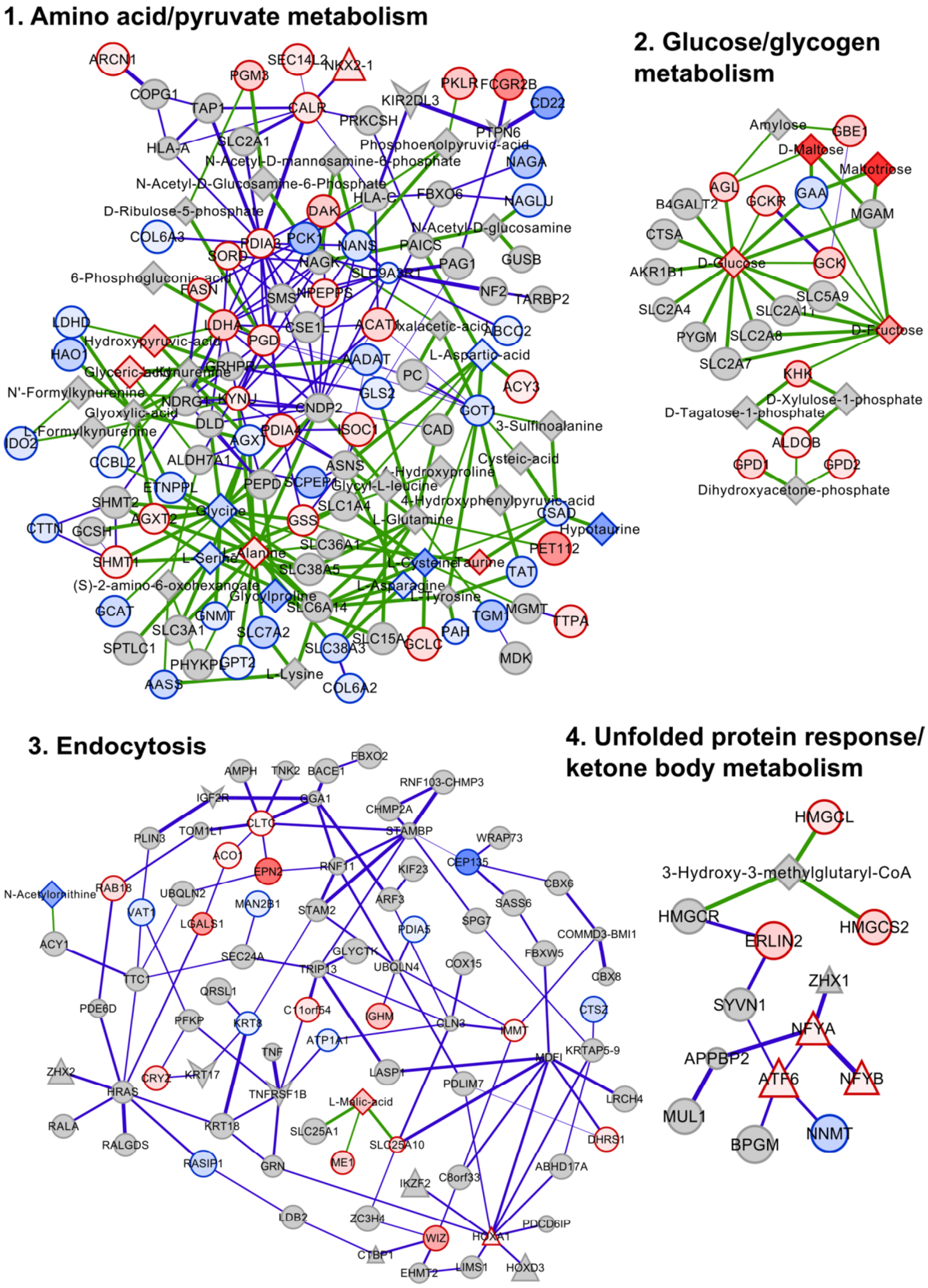


Figure 2-S7. PCSF model subnetworks.

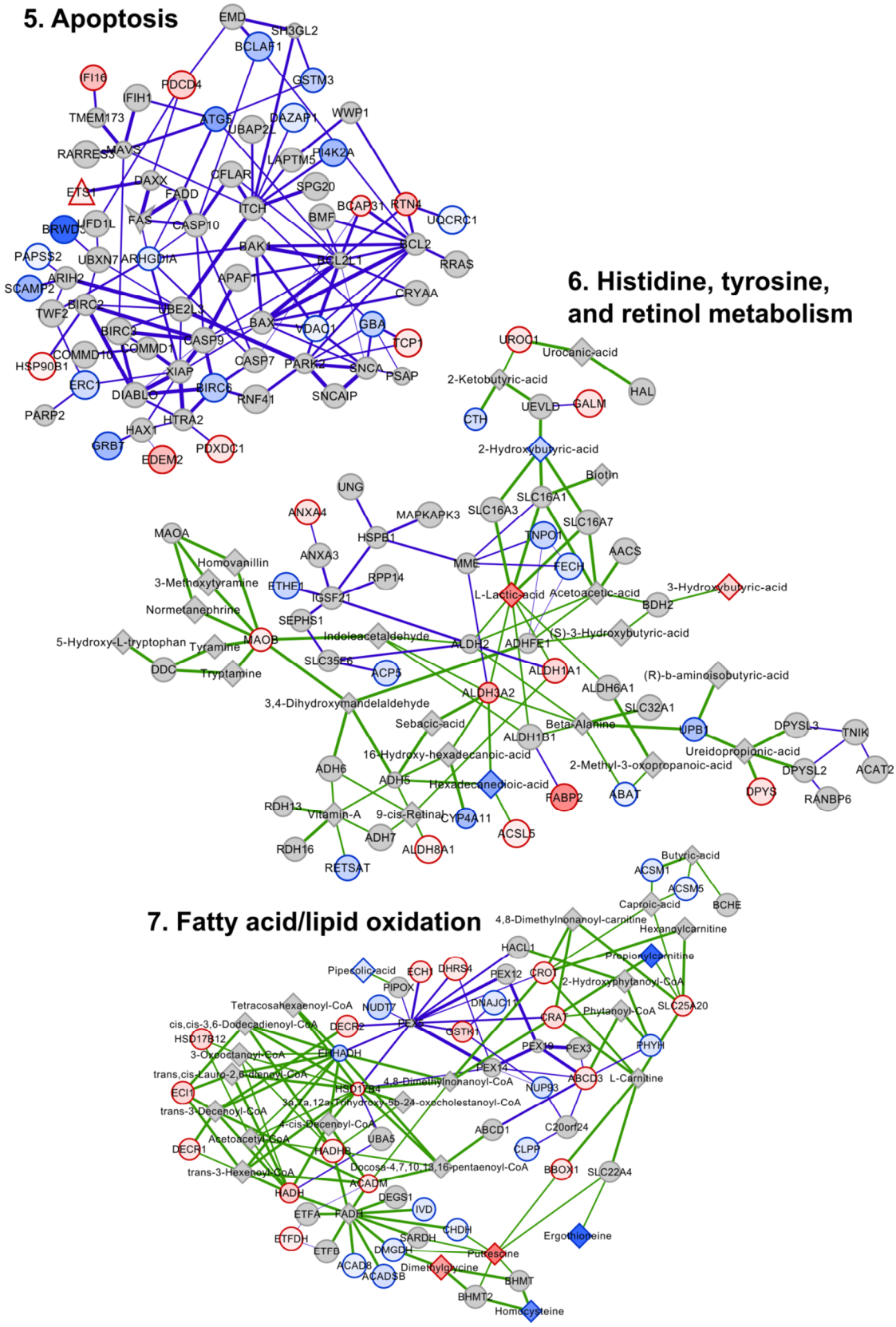


Figure 2-S7 (continued). PCSF model subnetworks.

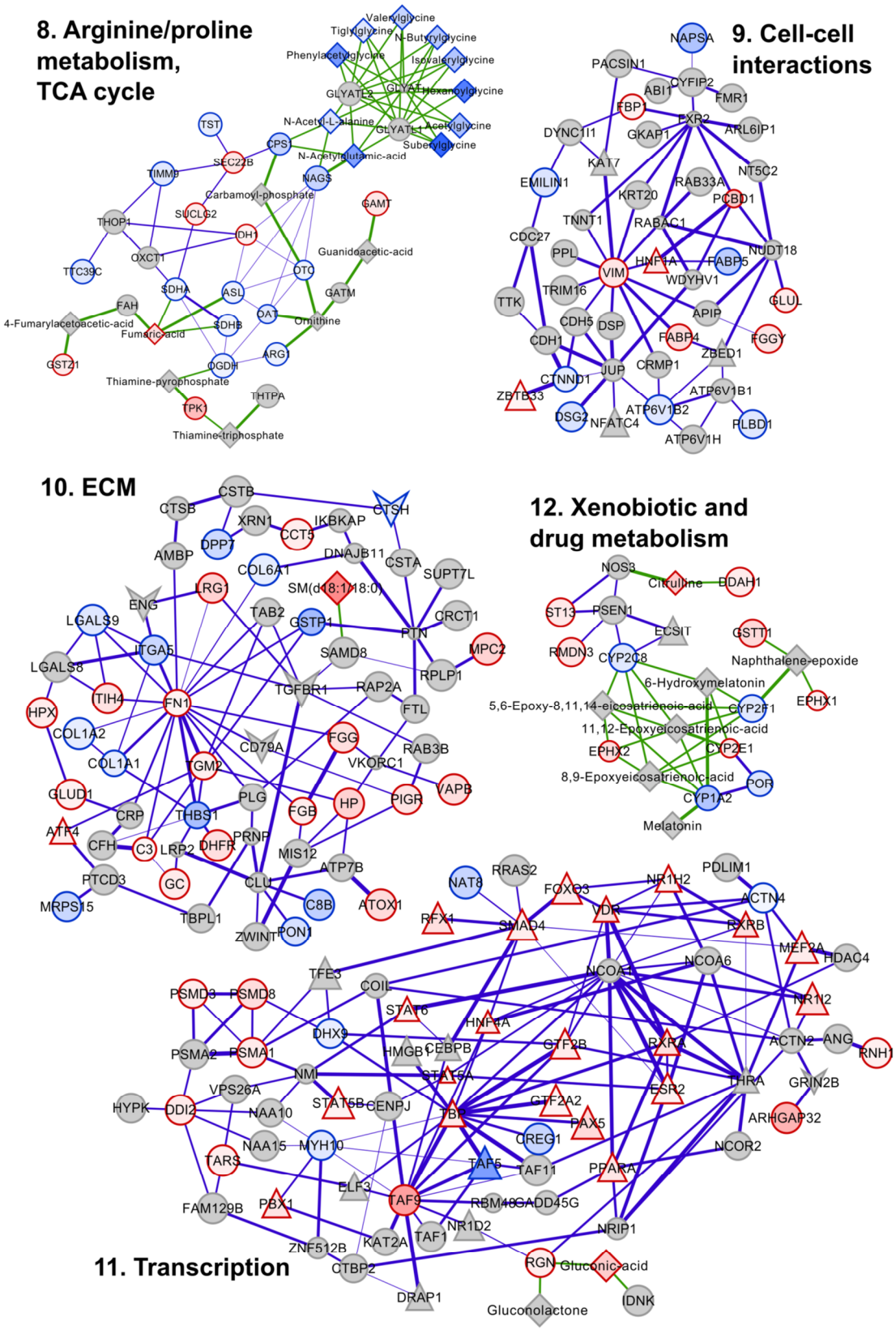


Figure 2-S7 (continued). PCSF model subnetworks.

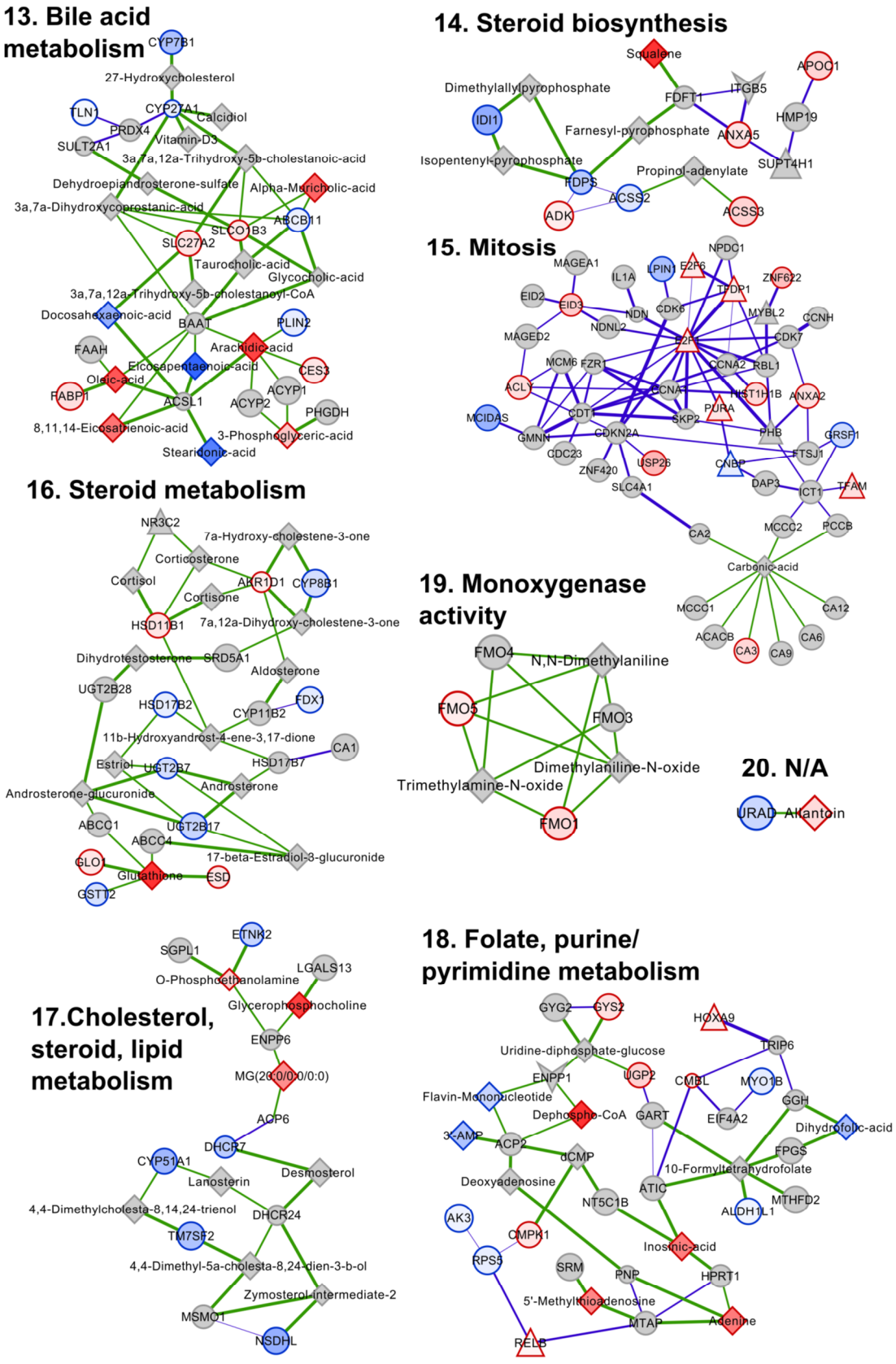


Figure 2-S7 (continued). PCSF model subnetworks.

CHAPTER 3

TEMPORAL TRANSCRIPTIONAL PROFILING REVEALS RGS4 AS A SELECTIVE MEDIATOR OF HEPATIC METABOLIC ADAPTATION TO OBESITY-INDUCED INSULIN RESISTANCE

Obesity promotes the development of insulin resistance, hyperglycemia, and metabolic syndrome which can lead to β-cell dysfunction and type 2 diabetes. In this study we profiled the hepatic transcriptomes of mice fed a standard chow diet (CD) or a short- or long-term high-fat diet (HFD). We found that diet progressively dysregulated the hepatic transcriptional landscapes of these mice. We also treated mice with the type 2 diabetes drug metformin and profiled their transcriptomes. We found that metformin only modestly affected transcriptional changes in these mice, although it significantly improved a number of physiological parameters related to insulin and glucose sensitivity. In addition, we performed temporal transcriptional profiling following insulin stimulation in CD and long-term HFD-fed mice. Insulin induced a robust transcriptional response in CD mice that was almost completely blunted in HFD. However, we observed a small set of 137 genes specifically modulated by insulin in HFD livers. Among these were regulators of G-protein signaling (RGS) genes, particularly *Rgs4*. We validated these findings and demonstrated that RGS4 protein expression is also up-regulated in HFD-fed mice following insulin stimulation. We fed liver-specific insulin receptor knock-out (LIRKO) mice a HFD and found that this effect of insulin on *Rgs4* gene expression was eliminated by LIRKO. We additionally obtained *Rgs4* knock-out mice and found that these mice are more insulin resistant compared to wild-type controls following HFD. Thus, we found a potentially novel mechanism by which hepatic insulin signaling is partially maintained following HFD.

This chapter presents work that is currently being prepared as a manuscript for submission.

Anthony R. Soltis[*], Norman J. Kennedy[*], Xiaofeng Xin, Santiago Vernia, Randall Friedline, Yoon Sing Yap, Bryan J. Matthews, Forest M. White, Douglas A. Lauffenburger, Jason K. Kim, Ernest Fraenkel[Δ], and Roger J. Davis[Δ]

*Denotes equal contribution. [Δ]Denotes equal contribution.

Author contributions: Conceptualization and methodology, A.R.S., N.J.K., D.A.L., F.M.W., J.K.K., R.J.D., and E.F.; Software and formal analysis, A.R.S.; Investigation, N.J.K., X.X., S.V., R.F., Y.S.Y., and B.J.M.; Resources, J.K.K., R.J.D., and E.F.; Writing, A.R.S. and N.J.K.; Visualization, A.R.S. and N.J.K.; Supervision, E.F., R.J.D., J.K.K., F.M.W., and D.A.L.; Project administration, R.J.D.

Acknowledgments: We thank Vicky Benoit for technical assistance and Kathy Gemme for administrative assistance. We acknowledge members of the MIT BioMicro Center for assistance with sequencing data collection. The Mouse Metabolic Phenotyping Center at the University of Massachusetts is supported by grant DK093000. This work was supported by R24 DK-090963 (E.F., R.J.D., F.M.W., and D.A.L.), R01 NS-089076 (E.F.), R01 DK107220 (R.J.D.), and used computing resources funded by the National Science Foundation under Award No. DB1-0821391 and sequencing support from NIH (P30-ES002109). R.J.D. is an investigator of the Howard Hughes Medical Institute.

Accession numbers: The mRNA-Seq raw and processed data are deposited in the Gene Expression Omnibus (GEO) as accession numbers GSE92488 and GSE77625.

3.1. INTRODUCTION

Human obesity is a major world-wide health problem that promotes hyperglycemia, insulin resistance, and, ultimately, type 2 diabetes [5, 6, 8]. During obesity, increased adipose tissue mass enhances the release of free fatty acids, along with hormones and proinflammatory cytokines [8]. These free fatty acids are effective signaling molecules whose accumulation in muscle and liver is strongly associated with insulin resistance [13].

Proper liver function is critical to maintaining normal metabolic health. Insulin regulates glucose homeostasis by increasing its uptake in peripheral tissues (primarily skeletal muscle) and via inhibition of hepatic gluconeogenesis [28]. In the liver, insulin binds and activates the insulin receptor and initiates signaling cascades that suppress gluconeogenesis and promote glycogen synthesis [33]. Elimination of insulin signaling in hepatocytes by insulin receptor knock-out (LIRKO) causes insulin resistance and hepatic dysfunction [41]. Accumulation of lipid metabolites in the liver activates pathways that disrupt normal insulin signaling, including PKCε activation and subsequent inhibition of the insulin receptor's kinase activities [35]. Such mechanisms that disrupt the liver's ability to suppress glucose production, however, are still not fully understood. For instance, mice lacking *Akt1*, *Akt2*, and *Foxo1*, three genes encoding proteins involved in the canonical insulin signaling pathway, are still able to suppress hepatic gluconeogenesis in response to insulin [45]. Therefore, analysis of hepatic insulin resistance can reveal new molecular mechanisms that may be exploited for therapeutic benefit against diseases like type 2 diabetes.

To study obesity-induced hepatic insulin resistance, we fed mice short (6 week) or long-term (16 week) high-fat diets (HFD) and compared these to normal chow diet (CD) fed controls. We used RNA-Seq to comprehensively profile hepatic transcriptional responses to these diets. We additionally treated CD and HFD-fed mice with metformin, the most common type 2 diabetes drug that principally acts in the liver to suppress hepatic glucose production [24, 31], to profile its effects on hepatic transcription. We also analyzed the effects of insulin stimulation in CD and long-term HFD-fed mice, performing temporal transcriptional profiling on livers treated with the hormone for variable lengths of time. This allowed us to compare transcriptional responses to insulin in normal and obese livers directly.

Our analyses demonstrated that both short and long-term HFD induce extensive changes in hepatic gene expression, that metformin induces modest effects on these responses, and that long-term HFD almost completely eliminates normal transcriptional responses to insulin. We did find, however, a small set of 137 genes that uniquely respond to insulin in HFD livers. Among these are genes that encode regulators of G-protein signaling. In particular, we found that *Rgs4* is

specifically induced by insulin stimulation in the hepatocytes of HFD-fed mice and that this effect is dependent on the presence of the insulin receptor. We further characterized the role of this gene by feeding liver-specific *Rgs4* knock-out mice a HFD. We found that *Rgs4*-null mice are more insulin resistant on a HFD compared to wild-type controls. Thus, we identified a candidate gene whose expression is uniquely insulin-sensitive in HFD livers that appears to play a role in conferring insulin sensitivity following this diet.

3.2. RESULTS

3.2.1. HFD feeding progressively degrades metabolic health and promotes obesity and insulin resistance, while metformin improves health during HFD

We examined diet-induced obesity in mice by feeding a HFD for 6 or 16 weeks. We also fed control mice a standard laboratory CD. We examined several physiological parameters to confirm that feeding a HFD causes obesity and insulin resistance. Glucose and insulin tolerances tests demonstrated that both 6 and 16 week HFD-fed mice were severely intolerant to both insulin and glucose compared to CD-fed controls (**Figure 3-S1A-B**). Additionally, immunoblots of insulin-stimulated AKT activation in these mice showed progressive decreases for this molecular readout as HFD feeding duration increased (**Figure 3-S1C-E**). We also performed hyperinsulinemic-euglycemic clamp studies to directly assess insulin resistance in conscious mice across all conditions. HFD feeding caused a progressive decrease in glucose infusion rate (a measurement of whole body insulin sensitivity), increased hepatic glucose production, and decreased hepatic insulin action, whole body glucose turnover, glycolysis, and glycogen plus lipid synthesis during the clamps (**Figure 3-1A**). We also found that glucose uptake by the gastrocnemius muscle and epididymal adipose tissue during the clamps was progressively suppressed by HFD feeding (**Figure 3-S2A-B**). These data confirm that consumption of a HFD causes progressively severe insulin resistance.

In addition to dietary perturbations, we also treated groups of CD, 6 week, and 16 week HFD-fed mice with the widely prescribed type 2 diabetes drug metformin. During hyperinsulinemic-euglycemic clamps, we found that metformin treatment in 16 week HFD-fed mice improved

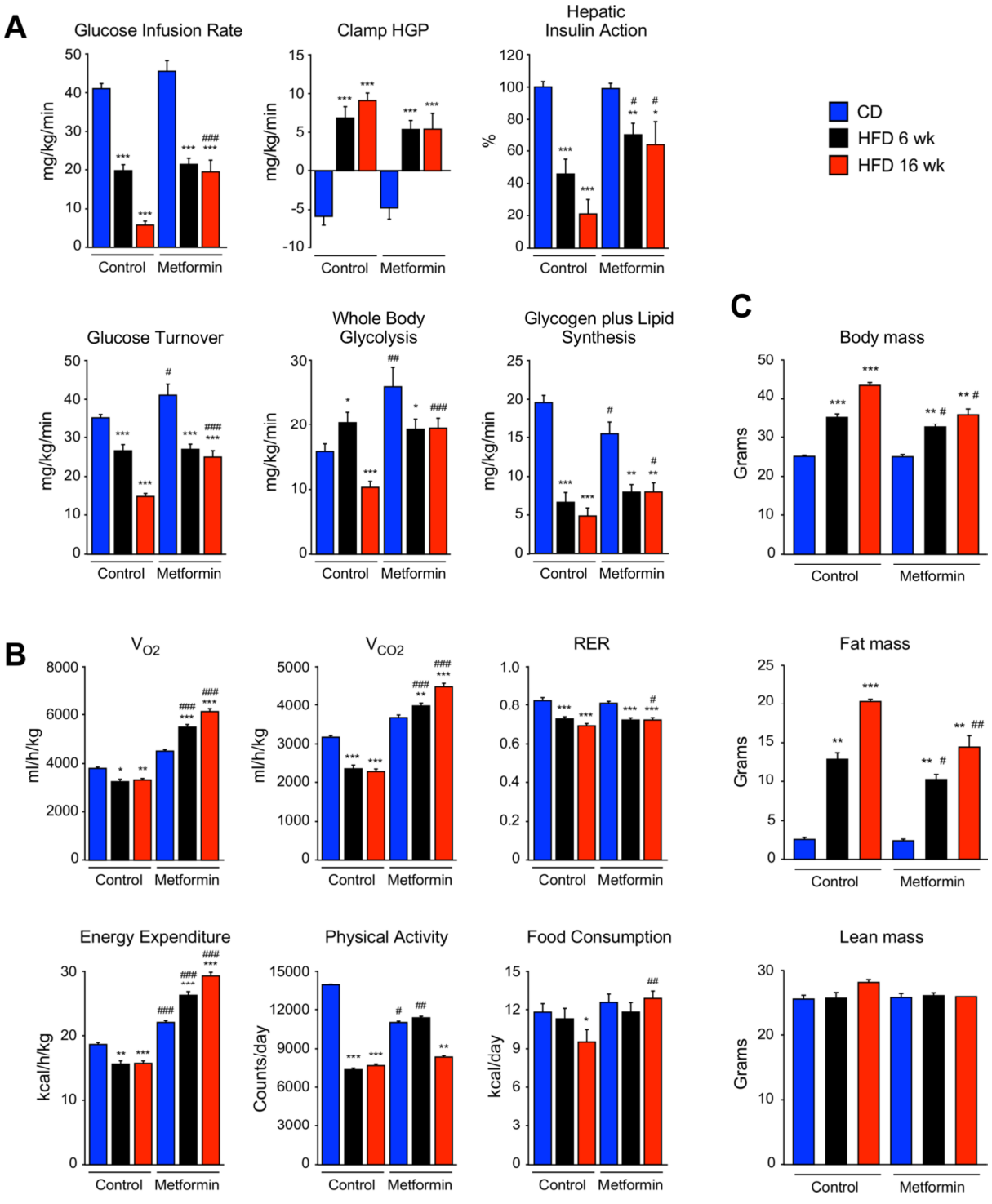


Figure 3-1. Physiological analysis of HFD-fed mice.

Figure 3-1 (continued). Physiological analysis of HFD-fed mice. (A) CD-fed and HFD-fed (6 weeks and 16 weeks) mice at age 24 weeks were treated without or with metformin and examined using hyperinsulinemic-euglycemic clamps to measure the glucose infusion rate, clamp hepatic glucose production (HGP), hepatic insulin action, whole body glucose turnover, whole body glycolysis, and glycogen plus lipid synthesis (mean ± SEM; n = 8). (B) Mice were examined using metabolic cages to measure V_{O2}, V_{CO2}, respiratory exchange ratio (RER), energy expenditure, physical activity, and food consumption (mean ± SEM; n = 8). (C) Mice were examined using ^{1}H-MRS to measure total body mass, fat mass, and lean mass (mean ± SEM; n = 8). * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$ vs CD. # $p < 0.05$; ## $p < 0.01$; ### $p < 0.001$ vs Control HFD.

several measures of metabolic health, including enhanced glucose infusion rates, insulin action, and whole body glycolysis (**Figure 3-1A**). Metformin treatment in 6 week HFD-fed mice produced minimal effects on these same parameters, where mice only showed improved insulin action compared to non-treated 6 week HFD-fed controls.

We also used metabolic cage analyses to further characterize the metabolic states of these mice. We found that both 6 and 16 week HFD caused decreased consumption of O_2 (V_{O2}), release of CO_2 (V_{CO2}), respiratory exchange ratio, energy expenditure, and physical activity (**Figure 3-1B**). Such changes contribute to the development of obesity. Indeed, ^{1}H-MRS analysis demonstrated that HFD feeding caused a progressive increase in fat mass in the absence of significant changes in lean tissue mass (**Figure 3-1C**). With metformin treatment, we observed increases in V_{O2}, V_{CO2}, energy expenditure, and physical activity (**Figure 3-1B**) coupled with reduced fat mass in both 6 and 16 week HFD-fed mice (**Figure 3-1C**). Together, these data establish that feeding a HFD causes obesity and insulin resistance that partially develops within 6 weeks and fully develops after 16 weeks. We also demonstrated that metformin improves many characteristics associated with metabolic health, particularly in mice fed long-term HFD.

3.2.2. HFD induces extensive and progressive hepatic transcriptional dysregulation while metformin drives modest alterations to hepatic transcriptomes

We used RNA-Seq to profile the hepatic transcriptomes of CD, 6 week, and 16 week HFD mice treated without or with metformin. Using non-metformin treated CD livers as controls, we

detected greater than 3,000 genes differentially expressed in at least one of the five treatment conditions (**Figure 3-2A**). Diet by far played the greatest role in determining the extent of transcriptional dysregulation. Metformin treatment alone in CD livers induced expression changes in a set of 266 genes that are modestly enriched in glutathione transferase activities. HFD induced expression changes in 1,137 and 2,507 genes following 6 and 16 weeks of feeding, respectively. The majority of genes whose expression levels were altered by the shorter-term HFD are maintained at 16 weeks, where further exposure to HFD expands the total pool of affected genes (**Figure 3-2B**). Within the sets of genes commonly up-regulated by 6 and 16 weeks HFD (491 genes), we found strong enrichments for immune, stress, and defense responses, along with lipid metabolic processes and cytokine responses, while genes commonly down-regulated by both HFDs are involved in small molecule metabolic processes, including amino acids, carboxylic acids, and glutamine. Within the expanded set of genes altered by the longer term 16 week HFD are enrichments for extracellular and membrane components in the set of up-regulated genes, along with apical junction complex components in the down-regulated genes.

In HFD livers treated with metformin, we again detected more total expression changes following 16 week HFD feeding (1,882 genes) versus 6 weeks (1,452 genes). We found substantial overlap in the genes called differentially expressed in treated and non-treated HFD livers versus non-metformin treated CD controls (**Figure 3-2C**). In 16 week HFD livers, where we observed improved measures of insulin sensitivity as a consequence of treatment (**Figure 3-1**), metformin reduced the total pool of differentially regulated genes. Both treated and non-treated 16 week HFD livers showed expression changes in genes related to immune and defense responses, along with oxidation-reduction processes, lipid metabolism, and other small molecule metabolic processes. In genes uniquely differentially regulated in metformin treated livers, we found enrichments for sterol and cholesterol metabolic processes.

We also directly compared metformin treated HFD livers to their respective non-treated groups and, as observed when comparing CD treated and non-treated samples, found small total numbers of significantly differentially expressed genes (179 and 247 genes for 6 and 16 week HFD, respectively). We then asked which genes metformin consistently alters the expression of

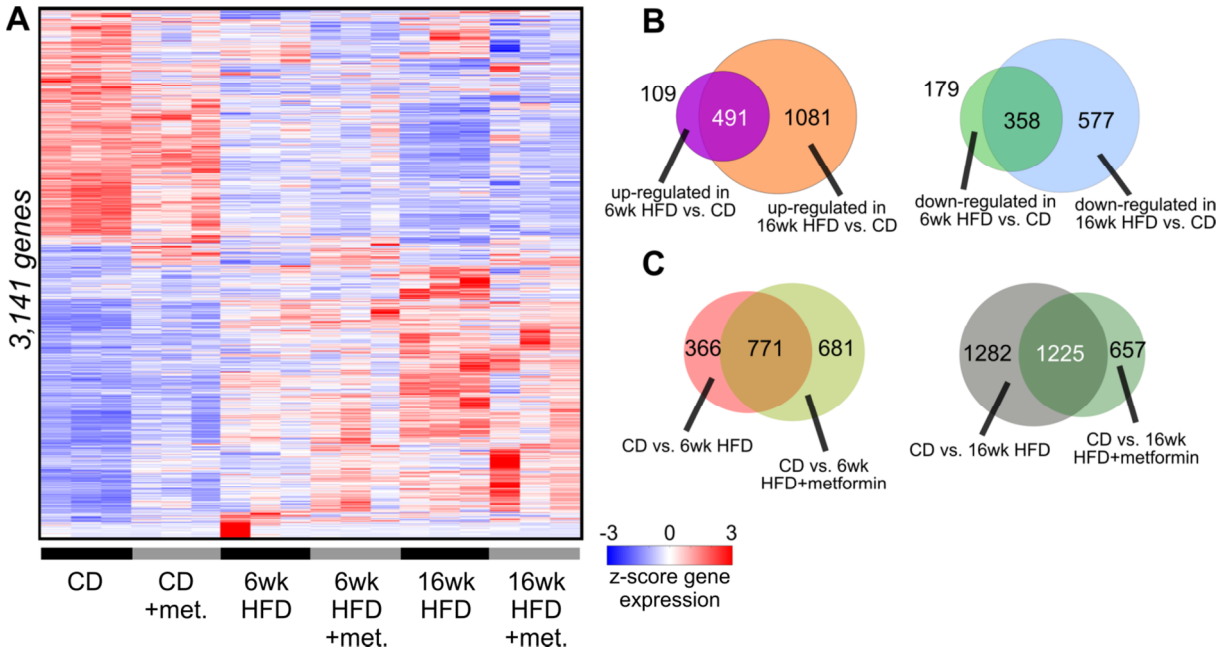


Figure 3-2. Short- and long-term HFD- and metformin-induced changes in liver mRNA transcription. (A) Heatmap of z-score normalized (gene-wise) mRNA expression measurements for 3,141 genes in CD, CD plus metformin (CD+met.), 6 week HFD, 6 week HFD plus metformin, 16 week HFD, and 16 week HFD plus metformin liver samples (n = 3 for all conditions). Displayed genes are those found to be differentially expressed ($|\log_2$ fold-change$| > 0.5$, q-value < 0.05) in any of five treatment conditions against control CD group. (B) Venn diagrams of genes found to be differentially expressed in 6 week HFD or 16 week HFD against CD. The left panel displays genes up-regulated in at least one of the HFD conditions while the right panel displays genes down-regulated by at least one of the HFD treatments. (C) The left panel shows a Venn diagram comparing genes found to be differentially expressed in 6 week HFD versus CD against 6 week HFD plus metformin versus CD. The right panel shows a Venn diagram comparing genes found to be differentially expressed in 16 week HFD versus CD against 16 week HFD plus metformin versus CD.

in all dietary conditions and found a small set of 24 such genes. Metformin consistently altered the expression levels of all these genes in the same direction, up-regulating 12, including *Acat3*, *Acot5*, *Cox7c*, and *Ring1*, and down-regulating 12, including *Cebpb*, *Cyp3a44*, and *Mbnl2*. Among these, *Cebpb* mRNA expression is known to be down-regulated by metformin in hepatocytes and deletion of this gene reduces hepatic steatosis and diabetes in *db/db* mice [242]. Thus, metformin appears to induce modest direct effects on hepatic transcription, though such changes may play critical roles in driving its therapeutic effects in the liver.

3.2.3. Insulin induces robust transcriptional responses in CD livers and a blunted, yet distinct, response in 16 week HFD samples

We next profiled hepatic responses to insulin by treating CD and 16 week HFD mice intraperitoneally with 1 U/kg of the hormone. We performed RNA-Seq on liver samples from mice stimulated for 15, 30, 60, and 120 minutes and compared each time point to their respective non-stimulated (PBS) baselines to identify insulin-sensitive genes. We uncovered 851 and 166 such genes in CD and 16 week HFD livers, respectively (**Figure 3-3A-B**). We identified a small set of 29 genes that are commonly insulin-sensitive in CD and 16 week HFD livers (**Figure 3-3C**), including *Sgk1* (up-regulated across both time courses), *Txnip* (down-regulated across both), *Fbf1* (up-regulated in CD, down-regulated in HFD), and *Hlx* (down-regulated in CD, up-regulated in HFD).

The majority of identified insulin-responsive genes are uniquely sensitive in mice fed a particular diet, with 822 genes specifically sensitive in CD livers and 137 genes sensitive in 16 week HFD livers alone. In CD livers, unique insulin-responsive genes are enriched for processes related to glucose homeostasis (e.g. up-regulated *Foxo1* and *Stat3* and down-regulated *Gyk* and *Hnf1a*), lipid biosynthesis (e.g. up-regulated *Cyp17a1*, *Insig1*, and *Ldlr* and down-regulated *Apoa4* and *Insig2*), and sequence-specific transcription factor activities (e.g. up-regulated *Atf3*, *Atf4*, *Fos*, *Foxo3*, *Jun*, *Rara*, and *Smad2* and down-regulated *Foxa1*, *Hhex*, *Ppara, Smad7*, and *Smad9*). In addition, insulin induced up-regulation of glycolytic enzymes, including *Gck* and *Pklr*, a response that is anticipated given insulin's known role in suppressing hepatic gluconeogenesis and promoting glycolysis. We also observed down-regulation of *Pdk4* by insulin in CD livers, which is also a known glycolysis-promoting hepatic response to insulin [243].

We used affinity propagation [244] to cluster the temporal expression profiles of insulin-responsive genes in CD livers (**Figure 3-3D, blue curves**). In general, expression changes occur early and return near baseline at later time points, occur early and remain altered throughout, or

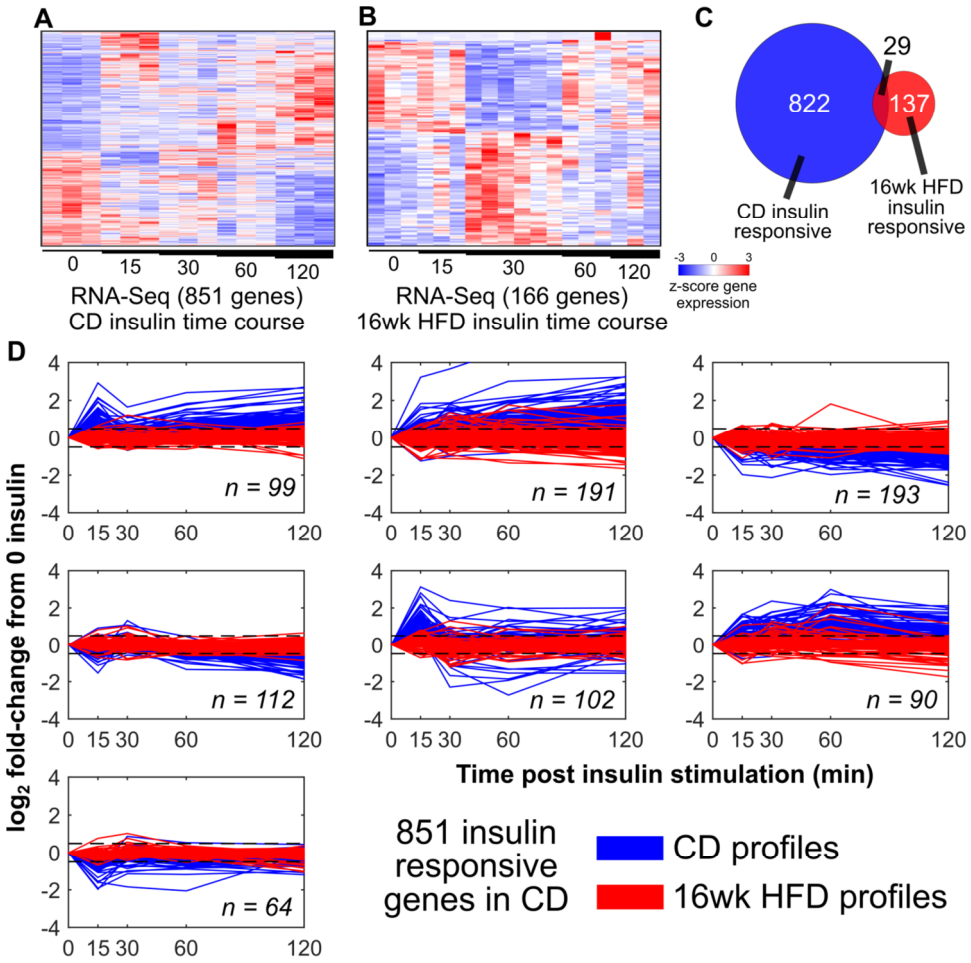


Figure 3-3. Temporal transcriptomic analysis following insulin stimulation in CD and 16 week HFD-fed mouse livers. (A) Heatmap of z-score normalized (gene-wise) mRNA expression measurements for 851 genes in CD livers found to be differentially expressed at any time point (15, 30, 60, and 120 minutes) post insulin stimulation compared to no insulin baseline control ($|\log_2$ fold-change$| > 0.5$, q-value < 0.05, n = 3 for all time points). (B) Heatmap of z-score normalized (gene-wise) mRNA expression measurements for 166 genes in 16 week HFD livers found to be differentially expressed at any time point (15, 30, 60, and 120 minutes) post insulin stimulation compared to no insulin baseline control ($|\log_2$ fold-change$| > 0.5$, q-value < 0.05, n = 3 for 0, 15, 60, and 120 minutes insulin, n = 6 for 30 minutes). (C) Venn diagram comparing insulin-responsive genes in CD and 16 week HFD insulin stimulation time courses. (D) Temporal transcriptional profiles of 851 insulin responsive genes in CD livers (blue traces) clustered into groups by affinity propagation (self-similarity = -7). Corresponding profiles for the same genes in 16 week HFD livers are shown as red traces in each plot. The numbers within each plot report how many CD insulin responsive genes fall into each cluster.

progressively increase or decrease across the time course. We then overlaid the 16 week HFD temporal profiles for these same genes onto the CD profiles to compare their temporal expression patterns, finding that these genes are indeed not responsive to insulin stimulation in HFD livers (**Figure 3-3D, red curves**). Thus, long-term HFD generally suppresses the normal transcriptional responses to insulin in the liver, consistent with the severe insulin resistant states of these mice.

3.2.4. Temporal analysis of insulin-induced transcription reveals *Rgs4* as an insulin-sensitive target in insulin resistant livers

While we observed that long-term HFD generally suppresses normal transcriptional responses to insulin, we did indeed observe expression changes in 166 genes following insulin stimulation in these livers, 137 of which are not significantly insulin-sensitive in CD livers (**Figure 3-4A**). Nearly all expression changes occurred at the 30 minute post-insulin stimulation time point and returned to near basal levels beyond this time. To enhance our confidence in these findings, we performed our 16 week HFD temporal analysis using additional mouse livers collected following 30 minutes of insulin treatment (for six total livers). Among these 137 genes, we observed four that are up-regulated by insulin stimulation and that encode proteins involved in termination of G-protein coupled receptor signaling, namely *Adrkb2* (or *Grk3*, G-protein-coupled receptor kinase 3), *Rgs1*, *Rgs2*, and *Rgs4* (regulators of G-protein signaling, or RGS). *Adrkb2* encodes a β_2-adrenergic receptor kinase that phosphorylates ligand-occupied receptors, thereby blocking signaling [245], and RGS proteins are GTPase activating proteins that inactivate G-proteins directly to shorten signaling [246].

We particularly focused additional analyses on *Rgs4* expression following insulin stimulation. It is established that RGS4 protein binds and inhibits the heterotrimeric GTPases $G\alpha_q$ and $G\alpha_{11}$ [247]. *Rgs4* gene expression specifically increased following hormone stimulation in 16 week HFD livers, more than doubling (2.25 fold-change) 30 minutes post stimulation before returning to basal levels at the end of the time course (**Figure 3-4B**). This *Rgs4* transcriptional response was not observed in CD livers treated with insulin. To confirm our RNA-Seq results, we measured hepatic *Rgs4* mRNA by TaqMan® assays. These experiments confirmed that *Rgs4*

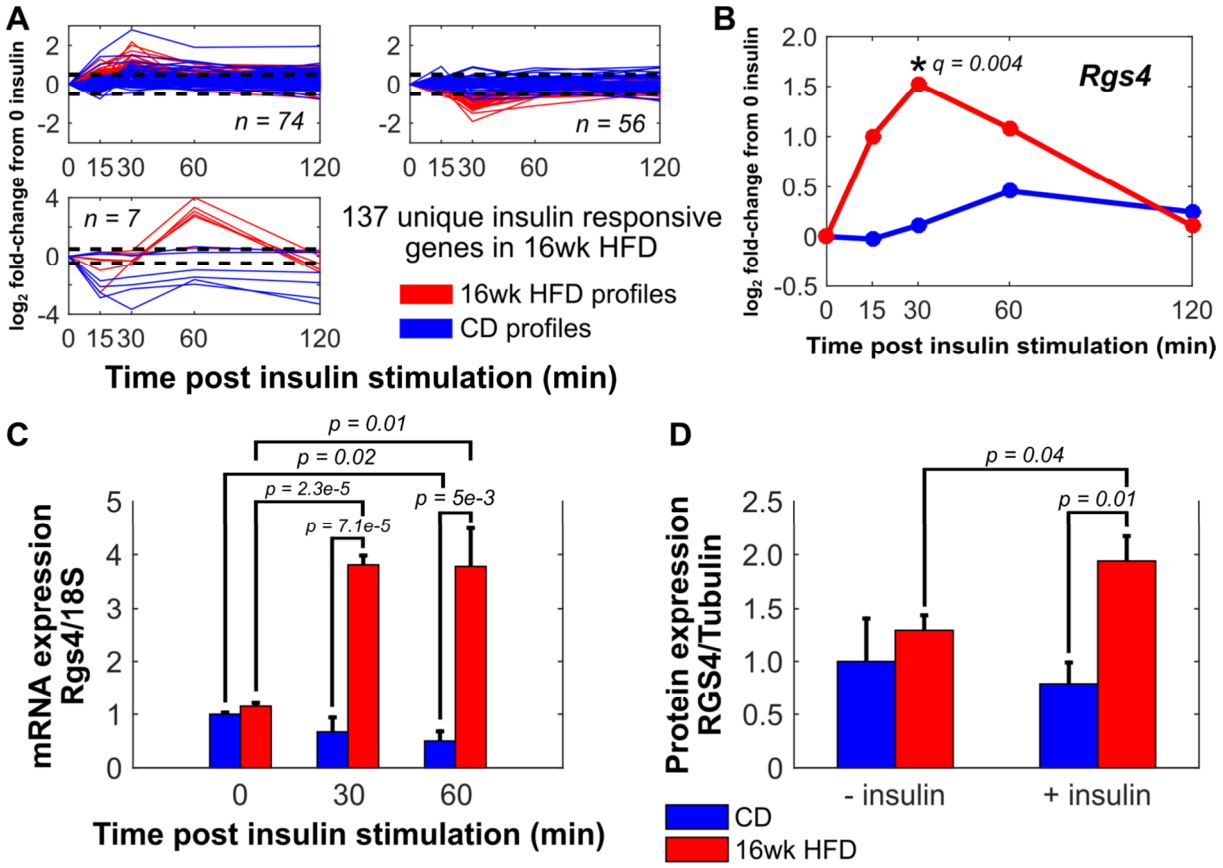


Figure 3-4. Insulin-stimulated transcriptional responses in 16 week HFD and Rgs4 mRNA and protein expression in liver. (A) Temporal transcriptional profiles of 137 genes uniquely insulin responsive in 16 week HFD livers (red traces) clustered into groups by affinity propagation (self-similarity = -3). Corresponding profiles for the same genes in CD livers are shown as blue traces in each plot. The numbers within each plot report how many 16 week HFD insulin responsive genes fall into each cluster. (B) Temporal transcriptional profiles for *Rgs4* gene in CD (blue) and 16 week HFD (red) livers. Marking at 30 minute time point in 16 week HFD represents significant differential expression for *Rgs4* against basal no insulin samples. (C) TaqMan mRNA expression measurements for *Rgs4* gene normalized to 18S mRNA in basal mouse livers or mice treated with 1 U/kg insulin for 30 or 60 minutes. P-values are reported for significantly differential comparisons. (D) Western immunoblot quantification for an antibody against RGS4 protein normalized to tubulin in basal CD and 16 week HFD mouse liver samples (-insulin) or in mice treated for 8 hours with 1 U/kg insulin (+insulin). P-values are reported for significantly differential comparisons.

expression is indeed increased in 16 week HFD livers following 30 minutes of insulin stimulation (**Figure 3-4C**). We also observed significant up-regulation of *Rgs4* mRNA transcript levels 60 minutes post-stimulation in HFD livers, along with significant down-regulation of *Rgs4*

in CD livers at this same time point. We also detected via Western immunoblots that RGS4 protein expression is indeed increased in 16 week HFD livers following long-term (8 hour) insulin stimulation. Together, these data indicate that insulin resistance generally does suppress normal insulin signaling, but that altered actions of insulin on liver regulatory pathways exist during insulin resistance. Here, we demonstrate that RGS4 is selectively altered by insulin in long-term HFD livers.

3.2.5. RGS4 is expressed in hepatocytes and its up-regulation by insulin in HFD livers is insulin receptor dependent

To confirm that RGS4 is indeed expressed and regulated in hepatocytes, we prepared primary hepatocytes from CD mouse livers and tested the effect of insulin on *Rgs4* mRNA expression in these cells. This analysis demonstrated that *Rgs4* is indeed expressed in hepatocytes and that the slight insulin-mediated inhibition of *Rgs4* mRNA expression we observed in whole CD livers is also observed in cultured primary hepatocytes (**Figure 3-5A**). We also found that treatment of primary hepatocytes with the inflammatory cytokine TNFα caused increased expression of *Rgs4* mRNA and protein (**Figure 3-5B-C**).

Given these results, we next sought to characterize whether or not the insulin-stimulated increase in *Rgs4* mRNA expression in HFD mice reflects signaling via the insulin receptor in hepatocytes. We obtained mice that are selectively insulin receptor deficient in hepatocytes (*Alb-cre*$^{-/+}$ *Insr*$^{LoxP/LoxP}$ or LIRKO) along with control mice (*Alb-cre*$^{-/+}$ *Insr*$^{+/+}$) (**Figure 3-S4**). We fed groups of LIRKO and control mice a CD or a 16 week HFD and measured *Rgs4* mRNA expression via RT-PCR following treatment with or without insulin. We found that insulin-stimulated increases in *Rgs4* gene expression were detected in HFD-fed control mice, but not in HFD-fed LIRKO mice (**Figure 3-5D**). Insulin receptor deletion did not affect *Rgs4* expression patterns in CD-fed mice. These results further demonstrate that the enhancement of *Rgs4* mRNA expression by insulin in HFD mice occurs in hepatocytes and that this effect is dependent on signaling via the hepatic insulin receptor. Thus, we provide further evidence that selective insulin receptor mediated signaling events are active in insulin resistant HFD-fed mouse livers.

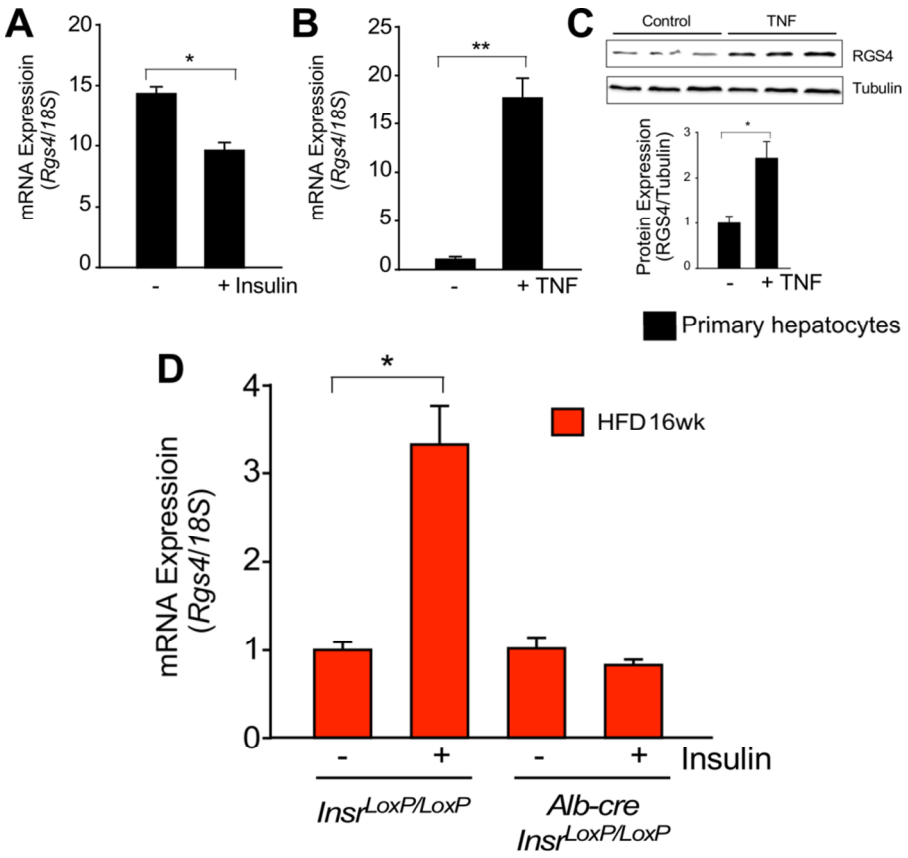


Figure 3-5. Insulin and TNFα regulation of Rgs4 in primary hepatocytes and Rgs4 expression in HFD-fed LIRKO mouse livers. (A, B) *Rgs4* mRNA expression was measured by RT-PCR in mouse primary hepatocytes following treatment without and with 100 nM insulin (30 minutes, A) and without and with 10 ng/ml TNFα (24 hours, B) (mean ± SEM, n = 3). (C) RGS4 protein expression was measured in mouse primary hepatocytes following treatment without and with 10 ng/ml TNFα for 24 hours (mean ± SEM, n = 3). (D) Hepatic *Rgs4* mRNA expression was measured by RT-PCR analysis in 16 week HFD-fed mice either expressing ($Insr^{LoxP/LoxP}$) or not expressing (*Alb-cre* $Insr^{LoxP/LoxP}$ or LIRKO) the liver insulin receptor treated without and with 1 U/kg insulin for 30 minutes. *$p < 0.01$; **$p < 0.001$.

3.2.6. RGS4 deletion in liver exacerbates HFD-induced insulin resistance

We demonstrated that RGS4 mRNA and protein expression are increased following insulin stimulation in 16 week HFD mice and that this up-regulation is dependent on the presence of the insulin receptor in hepatocytes. We next sought to characterize the significance of this action by insulin on RGS4 expression in the liver. It is established that RGS4 blocks $G\alpha_q$-mediated activation of PLCβ and subsequent activation of PKC [247]. This is significant because hepatic

PKC is activated by feeding a HFD (**Figure 3-S3**) and because this signaling pathway can cause hepatic insulin resistance [33]. Thus, insulin-stimulated RGS4 activation in HFD livers may serve to improve hepatic insulin sensitivity by limiting PKC signals. Elimination of this regulatory activity by RGS4 may further promote insulin resistance following HFD feeding.

To test the effect of RGS4 deletion in the liver, we obtained hepatocyte-specific deficient *Rgs4* mice (*Alb-cre*$^{-/+}$ *Rgs4*$^{LoxP/LoxP}$ or L-KO) and control wild-type (L-WT) mice (*Alb-cre*$^{-/+}$). We extracted DNA from the tails of both mice to check for the presence of the appropriate alleles (**Figure 3-6A**). Additionally, we extracted liver tissue DNA and performed PCR to identify the appropriate alleles (**Figure 3-6B**). We then confirmed that L-KO mice do not express RGS4 protein in the liver (**Figure 3-6C**). We fed both 8 week old L-WT and L-KO mice a HFD over a 16 week period and found that both sets of mice progressively gained weight, though L-KO mice gained slightly less weight than the L-WT controls at each time point (**Figure 3-6D**). At 16 weeks HFD, ^{1}H-MRS analysis showed no difference in lean mass between the two groups, though total mass and fat mass were slightly higher in WT mice compared to L-KO (**Figure 3-6E**). We performed glucose (**Figure 3-6F**) and insulin (**Figure 3-6G**) tolerance tests on 16 week HFD-fed L-WT and L-KO mice and found no effect of RGS4 deletion during the GTT. We did, however, observe a significant effect during the ITT where L-KO mice were more insulin resistant than controls. We additionally found that RGS4 deletion did not affect fasted blood glucose levels following HFD, but did observe lower levels of glucose in L-KO mice compared to L-WT controls (**Figure 3-6H**). These results indicate that RGS4 plays a role in regulating hepatic insulin sensitivity following HFD.

Additional experiments are currently underway to assess the role of RGS4 deletion on PKC activity in the liver. Hyperinsulinemic-euglycemic clamp studies will also be used to further characterize the effect of RGS4 deletion on insulin sensitivity and other relevant physiological parameters.

3.3. DISCUSSION

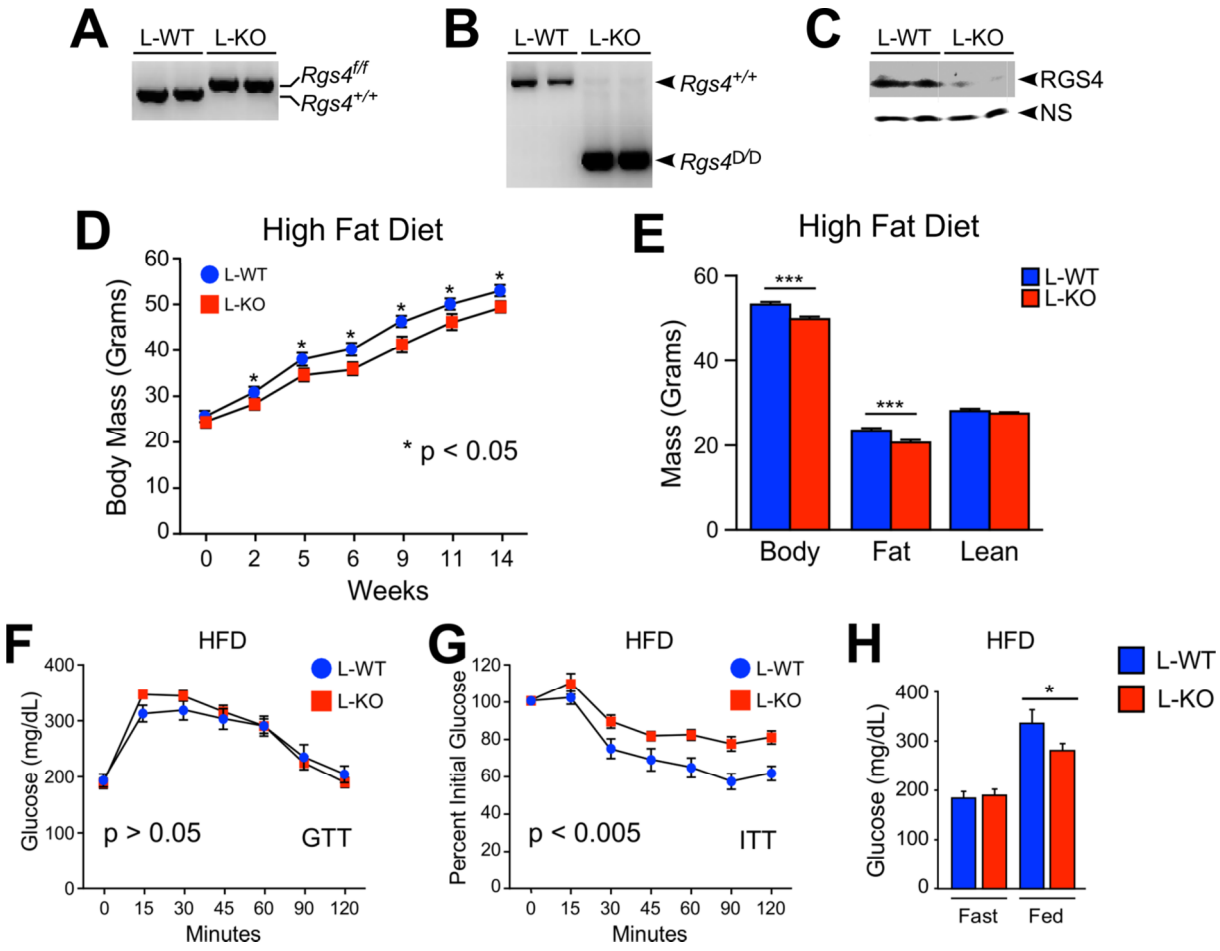


Figure 3-6. *Rgs4* deletion in mouse and physiological analysis. (A) DNA isolated from tail snips from control (L-WT) and liver *Rgs4* knockout mice (L-KO) was genotyped to identify wild-type and floxed *Rgs4* alleles. (B) DNA isolated from liver samples of floxed control (L-WT) and liver *Rgs4* knockout mice (L-KO) was analyzed by PCR to identify floxed and deleted *Rgs4* alleles. (C) Protein extracts prepared from L-WT and L-KO hepatocytes were analyzed for expression of RGS4 by immunoblot. (D) The body mass of L-WT and L-KO mice fed a HFD was examined over time (mean ± SEM; n=7-10, * $p < 0.05$). (E) L-WT and L-KO mice fed a HFD (16 weeks) were examined using ^{1}H-MRS to measure total body mass, fat mass, and lean mass (mean ± SEM; n = 9 (L-WT), n=10 (L-KO), *** $p < 0.005$). (F, G) HFD-fed (16 weeks) L-WT and L-KO mice were examined at age 24 weeks using glucose tolerance tests (GTT, F) and insulin tolerance tests (ITT, G) (mean ± SEM; n = 8-12). (H) Blood glucose levels of fed and fasted L-WT and L-KO mice fed a HFD (16 weeks) were examined (mean ± SEM; n=10). * $p < 0.05$.

In this study, we examined the hepatic transcriptomes of mice fed normal chow, short-term high-fat, and long-term high-fat diets. We used insulin and glucose tolerance tests, hyperinsulinemic-euglycemic clamps, and metabolic cage analyses to characterize the physiological effects of

HFD in these mice. These assays demonstrated that mice fed HFD progressively show signs of severe insulin resistance and glucose intolerance. We additionally treated these mice with the type 2 diabetes drug metformin. We found that metformin had a modest effect on transcription compared to HFD (in terms of raw numbers of affected genes), but highlighted a few key genes consistently altered by the drug across the diets, some of which are known targets of metformin action (e.g. *Cebpb*), that may be linked to the drug's mechanism of action in the liver [242]. Metformin did improve a number of physiological readouts related to insulin action in HFD livers. In addition, we performed temporal transcriptomic profiling of CD and 16 week HFD mouse livers following treatment with intraperitoneal insulin. We observed a robust transcriptional response to insulin in CD livers (>800 altered genes), but found that the vast majority of these insulin-dependent changes were blunted by HFD. We did observe, however, a small set of 137 genes that uniquely respond to insulin in HFD livers. Among these are genes encoding RGS proteins, particularly *Rgs4*.

We confirmed our finding of selective *Rgs4* insulin sensitivity with targeted gene expression analysis in mouse livers. We also used primary mouse hepatocytes to demonstrate that *Rgs4* is indeed expressed and sensitive to insulin and TNFα in this cell type. In addition, we showed that RGS4 protein expression is elevated following insulin stimulation in HFD mice. To test whether or not these results reflect signaling through the insulin receptor, we treated HFD-fed mice expressing or lacking the hepatic insulin receptor with insulin and found that the presence of this protein was necessary for the up-regulation of *Rgs4* gene expression by insulin during HFD. We additionally obtained mice either expressing (L-WT) or lacking (L-KO) liver *Rgs4* to further test the function of this gene. We found that L-WT and L-KO both gained weight on a HFD, though L-KO mice gained slightly less weight at each time point tested. This was due to changes in the amount of accumulated fat mass. We performed insulin and glucose tolerance tests on these mice and found no differences between the groups during GTT, but found a significant impairment of insulin sensitivity during ITT in L-KO mice compared to L-WT. Thus, RGS4 appears to play a role in regulating hepatic insulin sensitivity following HFD.

The lack of a robust transcriptional response to insulin in 16 week HFD mice was in itself not very surprising given that these mice are severely insulin resistant and glucose intolerant.

However, the finding of a distinct transcriptional response in HFD mice that, at least in the case of *Rgs4*, was dependent on the presences of the hepatic insulin receptor was intriguing. The observation of selective versus total hepatic insulin resistance in diabetic mice provides evidence towards this notion of active insulin signaling following HFD [40]. Selective insulin resistance is a state whereby insulin fails to suppress hepatic glucose production but still induces lipogenesis, creating a simultaneous hyperglycemic and hyperlipidemic state [39]. Total insulin resistance by LIRKO disrupts both branches of hepatic insulin action [40, 41]. More recent evidence proposes that liver insulin signaling is mostly intact and capable of promoting hepatic lipogenesis during type 2 diabetes, while enhanced intrahepatic FFAs, particularly acetyl CoA, resulting from enhanced lipolysis from white adipose tissue inhibit suppression of glucose production independent of insulin signaling [47, 48]. Our transcriptional data from HFD mice and results from HFD-fed LIRKO mice support this view of at least partially intact hepatic insulin signaling in obese mice.

RGS proteins are GTPase activators that inhibit and shorten signaling through G-proteins [246]. RGS4 in particular blocks $G\alpha_q$-mediated activation of PLCβ and subsequent activation of PKC [247]. This protein has mostly been studied in the context of neurological functions and diseases, including opiate tolerance and dependence [248], Parkinson's disease [249], and schizophrenia [250, 251], along with heart failure [252]. We showed here that PKC activation is higher in 16 week HFD-fed mice compared to CD controls, consistent with earlier reports of hepatic lipid accumulation leading to protein kinase C epsilon (PKCε) activation [33, 38]. PKC can associate with the hepatic insulin receptor and impair kinase signaling, while knock-down of PKCε in rat livers protects them from lipid-induced insulin resistance [38]. RGS4 activation in hepatocytes following HFD may act as a compensatory mechanism that limits signaling through the PKC pathway to enhance canonical insulin signaling (**Figure 3-7**). Indeed, we found that mice lacking RGS4 were more insulin resistant compared to wild-type controls when fed a long-term HFD. We did not observe a similar effect of RGS4 deletion during glucose tolerance tests. This could result from compensation via increased insulin release from the pancreas during GTT in HFD-fed L-KO mice. *We are currently undertaking additional experiments to assess the role of RGS4 deletion on PKC activity in the liver. These studies can directly indicate effects of RGS4 on downstream G-protein signaling. We are also performing hyperinsulinemic-euglycemic clamp*

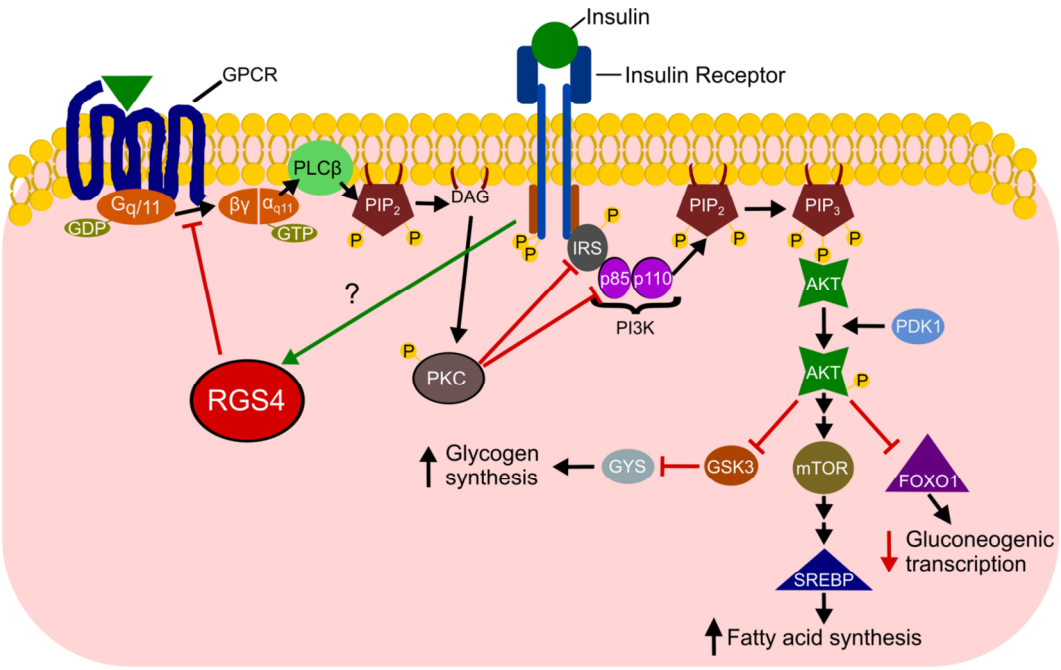


Figure 3-7. Proposed mechanism of RGS4 activity in HFD livers. Signaling through G-protein coupled receptors (GPCR) during HFD leads to up-regulation of protein kinase C (PKC) activity, which can inhibit up-stream components of the canonical insulin signaling pathway. RGS4 activity is increased via an unknown, yet insulin receptor-dependent, mechanism in HFD livers specifically. RGS4 is known to decrease signaling through GPCRs via its GTPase activities, thereby likely limiting some signaling potential toward downstream PKC activation. Abbreviations: insulin receptor substrates (IRS); phosphatidylinositol-3-OH kinase (PI3K); phosphatidylinositol (4, 5) bisphosphate (PIP_2); phosphatidylinositol (3, 4, 5) triphosphate (PIP_3); protein kinase B (AKT); 3-phosphoinositide dependent protein kinase-1 (PDK1); glycogen synthase kinase 3 (GSK3); glycogen synthase (GYS); mechanistic target of rapamycin (mTOR); sterol regulatory element binding protein (SREBP); forkhead box protein O1 (FOXO1); Heterotrimeric G_q protein ($G_{q/11}$); Guanosine di(tri)phosphate (GDP/GTP); G protein α_{11} subunit (α_{q11}); G protein βγ subunit (βγ); phospholipase C β (PLCβ); diacyl glycerol (DAG).

studies to further characterize the effect of RGS4 deletion on insulin sensitivity and other physiological parameters. Thus, in this work we identified a potentially novel mechanism by which insulin resistant livers preserve some level of canonical insulin signaling.

3.4. MATERIALS AND METHODS

3.4.1. Animals

C57BL/6J (stock number 000664) mice, B6.Cg-Tg(Alb-cre)21Mgn/J mice (stock number 003574) [253] mice, B6.Cg-*Rgs4*^{tm1Sdlk}/J and B6.129S4(FVB)-*Insr*^{tm1Khn}/J (stock number 006955) [254] were obtained from The Jackson Laboratories. All mice used for these studies were backcrossed to the C57BL6/J strain (ten generations) and housed in a facility accredited by the American Association for Laboratory Animal Care. Male mice were fed either a control diet (CD, Prolab Isopro RMH 3000, Purina) or a high fat diet (HFD, S3282, Bioserve) at age 8 weeks (for 16 weeks HFD) or age 18 weeks (for 6 weeks HFD). All mice were euthanized at age 24 weeks. Fat and lean masses were noninvasively measured using ^{1}H-MRS (Echo Medical Systems).

Treatment of mice with metformin (Sigma, PHR1084) was initiated at age 8 weeks (for 16 week time points) or 18 weeks (for 6 week time points). Metformin was dissolved in drinking water to attain a dose of 270 mg/kg/day lean body mass by dilution of a freshly prepared 10 mg/ml stock solution. Lean mass and fat mass were determined initially at 8 weeks of age, and subsequently every 2 weeks, by ^{1}H-MRS analysis, and the preparation of metformin-treated water was adjusted biweekly to maintain 270 mg/kg/day lean body mass. Control studies showed that mice do not consume less water when metformin is added and that each mouse drinks an average of 4 ml of water per day. Metformin treated water was replaced three times per week for the duration of the experiment. All mice were euthanized at age 24 weeks.

For insulin treatment studies, CD- and HFD-fed mice were starved overnight after which 1 U/kg of insulin (Novolin R, Novo Nordisk) diluted in PBS (or PBS alone) was administered by intraperitoneal injection. The mice were euthanized at various times following insulin administration, and the livers were frozen prior to removal using clamps cooled in liquid nitrogen. Frozen livers were pulverized into a powder using a CryoPREP impactor (Covaris) and aliquots of pulverized liver were prepared and used for subsequent analyses.

All experiments were carried out in accordance with guidelines for the use of laboratory animals and were approved by the Institutional Animal Care and Use Committees (IACUC) of the University of Massachusetts Medical School.

3.4.2. Genotype analysis

Genomic DNA was examined by PCR analysis to identify *Cre* recombinase using amplimers 5'-TTACTGACCGTACACCAAATTTGCCTGC -3' and 5'-CCTGGCAGCGATCGCTATTTTCCATGAGTG -3' (450 bp *Cre* DNA fragment) and 5'- GTT TTG TAA AGG GAG CCG AC-3' and 5'- CCT GAC TAC TGA GCC TGG TTT CTC-3' (224 bp control DNA fragment). Genotyping of $Insr^{LoxP}$ mice was performed using the amplimers 5'-GATGTGCACCCCATGTCTG-3' and 5'-CTGAATAGCTGAGACCACAG-3' to detect the $Insr^{+}$ (279 bp) and $Insr^{LoxP}$ (313 bp) alleles. Genotyping of $Rgs4^{LoxP}$ mice was performed using the amplimers 5'-GCT CAC CTT GGG AAG TAG CA-3' and 5'-CTG TGT TCG CAG GAA TCT GA-3' to detect the $Rgs4^{+}$ (352 bp) and $Rgs4^{LoxP}$ alleles (400 bp). Detection of deleted Rgs4 alleles in liver was determined by PCR analysis of genomic DNA using amplimers 5'-GCT CAC CTT GGG AAG TAG CA-3' and 5'-CTG GAC CAC ATT CCT TCA TTC A-3' to identify $Rgs4^{+}$ (2927 bp) and $Rgs4^{\Delta}$ (502 bp) alleles.

3.4.3. Hyperinsulinemic-euglycemic clamp studies.

Clamp studies were performed at the Mouse Metabolic Phenotyping Center at the University of Massachusetts Medical School. A 2-hr hyperinsulinemic-euglycemic clamp was conducted using overnight fasted conscious mice with a primed and continuous infusion of human insulin (150 mU/kg body weight priming followed by 2.5 mU/kg/min; Humulin, Eli Lilly), and 20% glucose was infused at variable rates to maintain euglycemia [255].

3.4.4. Metabolic cages

Food/water intake, energy expenditure, respiratory exchange ratios, and physical activity were measured using metabolic cages (TSE Systems) by the Mouse Metabolic Phenotyping Center at the University of Massachusetts Medical School. The mice were housed under controlled temperature and lighting with free access to food and water.

3.4.5. Glucose and insulin tolerance tests

Glucose and insulin tolerance tests were performed by intraperitoneal injection of mice with glucose (1g/kg) or insulin (1.0 U/kg) using methods described previously [224].

3.4.6. Blood analysis

Blood glucose was measured with an Ascensia Breeze 2 glucometer (Bayer). Insulin concentration in plasma was measured by multiplexed ELISA using the Luminex 200 system (Millipore). Plasma alanine transaminase (ALT) and aspartate aminotransferase (AST) activity were measured using the ALT and AST Reagent kit (Pointe Scientific) with a Tecan Infinite M1000 plate reader (Tecan). Plasma concentrations of triglycerides and cholesterol were determined by FPLC analysis by the Mouse Metabolic Phenotyping Center at the University of Cincinnati.

3.4.7. Primary hepatocytes

Primary hepatocytes were prepared from mice as previously described [256]. Briefly, a modified 2-step perfusion method using Liver Perfusion Media and Liver Digest Buffer (Invitrogen) was performed. Cells were seeded on plates pre-coated (1 h) with collagen I (BD Biosciences) in plating medium (DMEM, 4.5 g/L glucose) supplemented with 10% FBS, 0.2% BSA, 2 mM glutamine, 1 mM sodium pyruvate, 100 units/ml penicillin, 100 µg/ml streptomycin, 1 µM dexamethasone, and 1 nM insulin. After attachment (2 h), the medium was removed, and the hepatocytes were incubated (22 h) in maintenance medium (DMEM 1.0 g/L glucose) supplemented with 0.2% BSA, 2 mM glutamine, 1 mM sodium pyruvate, 100 units/ml penicillin, 100 µg/ml streptomycin, and 0.1 µM dexamethasone). When indicated, the hepatocytes were incubated with TNFα (10 ng/ml (R&D Systems) diluted in PBS containing 0.5% fat-free BSA (Sigma). or insulin (100 nM, Sigma).

3.4.8. Quantitative RT-PCR

Total RNA was extracted from the pulverized frozen mouse livers (RNeasy kit, Qiagen), converted to cDNA (High Capacity RT kit, Life Technologies) and quantitative PCR analysis of mRNA expression was performed using a Quantstudio PCR machine (Life Technologies) and TaqMan® assays for *Rgs4* (Mm00501389_m1). A duplex PCR was performed in a single well using TaqMan® assays to quantify the target and 18S mRNA (catalog number 4308329; Life Technologies). The data are presented as relative mRNA expression normalized to 18S mRNA.

3.4.9. Immunoblot analysis

Protein extracts from pulverized liver and cultured hepatocytes were prepared in Triton lysis buffer (20 mM Tris [pH 7.4], 1% Triton X-100, 10% glycerol, 137 mM NaCl, 2 mM EDTA, 25 mM β-glycerophosphate, 1 mM sodium orthovanadate, 1 mM phenylmethyl-sulfonyl fluoride, and 10 μg/ml each of aprotinin and leupeptin). Protein content was quantified by the Bradford method (Bio-Rad). Standard techniques were used to separate cell extracts (15-80 μg of protein) by SDS-PAGE for immunoblot analysis using antibodies from Santa Cruz (RGS4), Cell Signaling (phospho-PKC^{Thr514}, AKT, phospho-AKT^{Ser473}), Millipore (phospho-Tyrosine) and Sigma-Aldrich (α-Tubulin). The primary antibodies were detected by incubation with anti-mouse or anti-rabbit IgG conjugated to infrared dyes (IRDye®, LI-COR Biosciences). Immune complexes were detected using the Odyssey infrared imaging system (LI-COR Biosciences).

3.4.10. mRNA-Seq and analysis

We prepared mRNA-Seq libraries from all 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 [122] (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) [225]. We restricted TopHat to only align to known transcript splice junctions.

We performed all differential expression analyses with DESeq2 (version 1.0.18) [227]. Within the DESeq framework and prior to running the generalized linear model, we used the Bioconductor package conditional quantile normalization (CQN, version 1.6.0) [226] to remove systematic biases due to GC-content and gene length coverage. We used the values calculated by CQN as offsets in the DESeq GLM (as recommended in the DESeq2 manual). For temporal transcriptional profiling following insulin stimulation, we ran the CD and 16 week HFD differential expression analyses with multi-contrast GLMs, using the respective non-insulin-stimulated conditions as baselines and each subsequent time point as contrasts against this. For dietary and metformin treatment comparisons, we used single contrast GLMs of each test condition against non-metformin treated CD samples as a baseline. Throughout, 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). For temporal analyses, a gene passing these criteria in at least one comparison against baseline was considered "insulin sensitive."

3.4.11. Clustering and enrichment analyses

All hierarchical clustering analysis was done with the *clustergram* function in Matlab with Euclidean distance and average linkage. For enrichment analyses, we used custom Matlab code implementing the hypergeometric distribution for enrichment p-value calculations and used the Benjamini-Hochberg FDR procedure [257] to correct for multiple hypotheses. In general, an FDR < 0.1 was deemed significant. We used mouse gene ontology (GO) terms (GOC validation date January 30, 2014; downloaded February 5, 2014) for all enrichment analyses [258].

3.5. SUPPLEMENTARY FIGURES

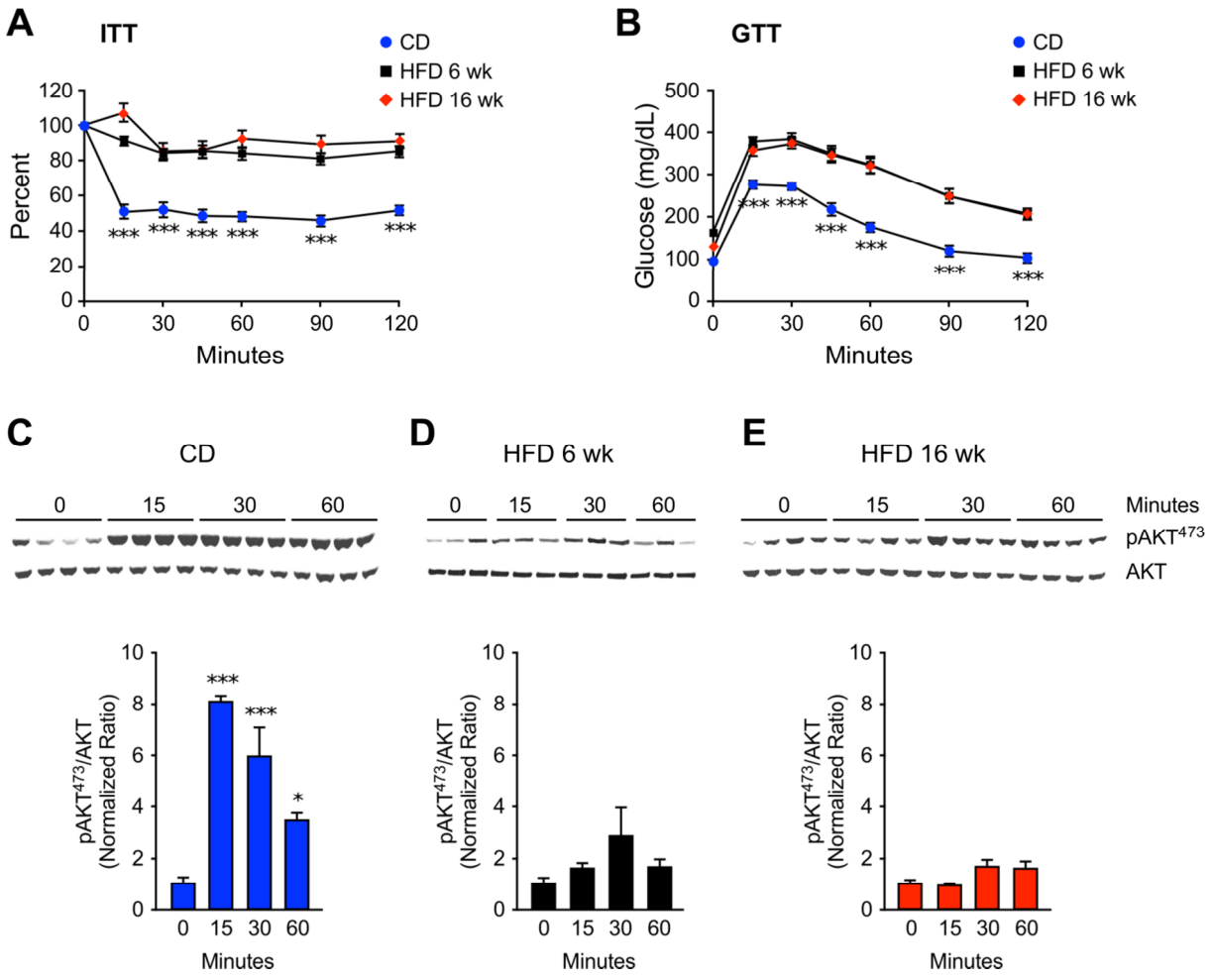


Figure 3-S1. Physiological analysis of HFD-fed mice. (A, B) CD-fed and HFD-fed (6 weeks and 16 weeks) mice were examined at age 24 weeks using insulin tolerance tests (ITT, A) and glucose tolerance tests (GTT, B) (mean ± SEM; n = 8-12). (C-E) CD-fed and HFD-fed (6 weeks and 16 weeks) mice were treated with insulin (1U/kg) by intraperitoneal injection. Hepatic AKT was examined by immunoblot with antibodies against pSer^{473}-AKT and AKT (mean ± SEM; n = 3~4).

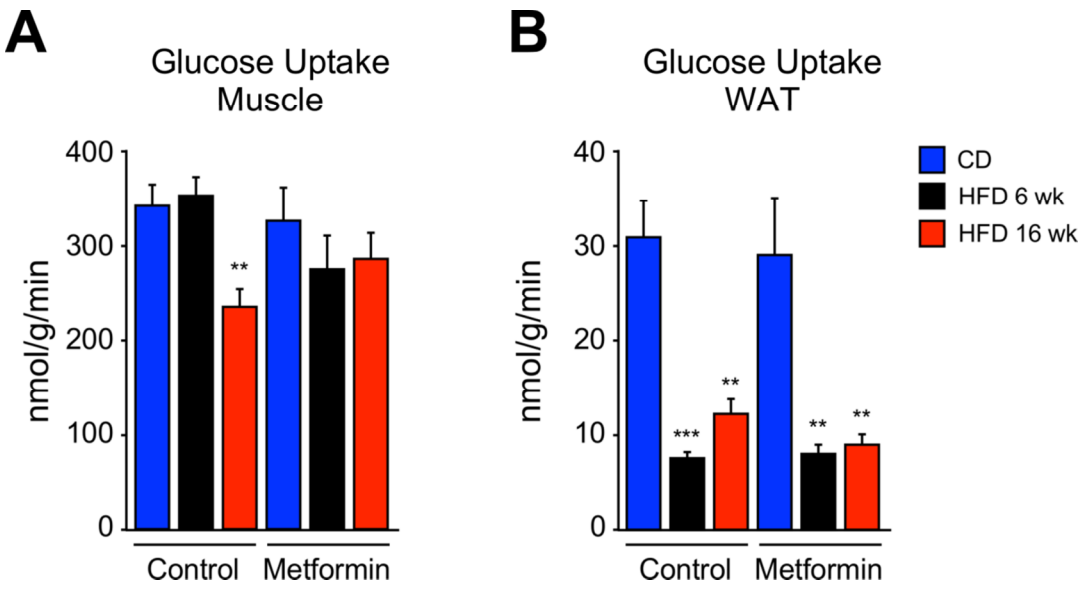


Figure 3-S2. Measurements of glucose uptake in vivo. (A, B) CD-fed mice and HFD-fed (6 weeks and 16 weeks) mice at age 24 weeks were treated without or with metformin and examined using the hyperinsulinemic-euglycemic clamp technique to measure glucose uptake by gastrocnemius muscle (A) and epididymal adipose tissue (B) (mean ± SEM; n = 8; **, p<0.01; ***, p<0.001).

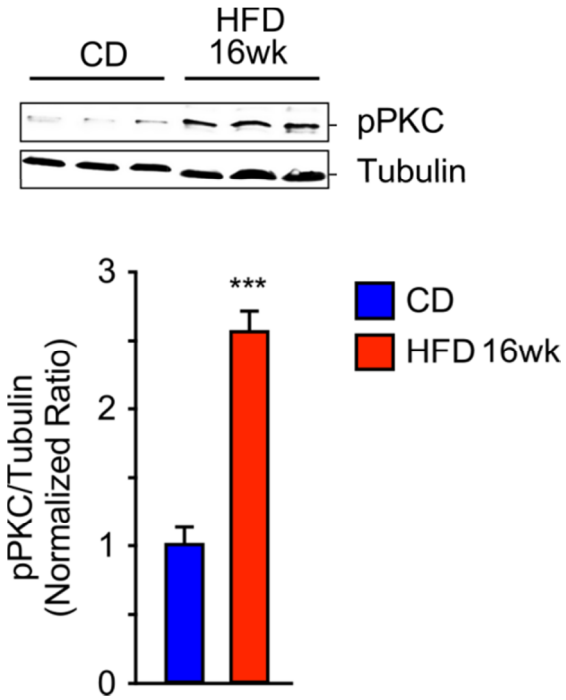


Figure 3-S3. Measurements of PKC activity in vivo. Hepatic extracts prepared from CD-fed mice and 16 week HFD-fed mice were examined by immunoblot analysis by probing with antibodies to phospho-PKC and Tubulin (mean ± SEM; n = 3; ***, p<0.001).

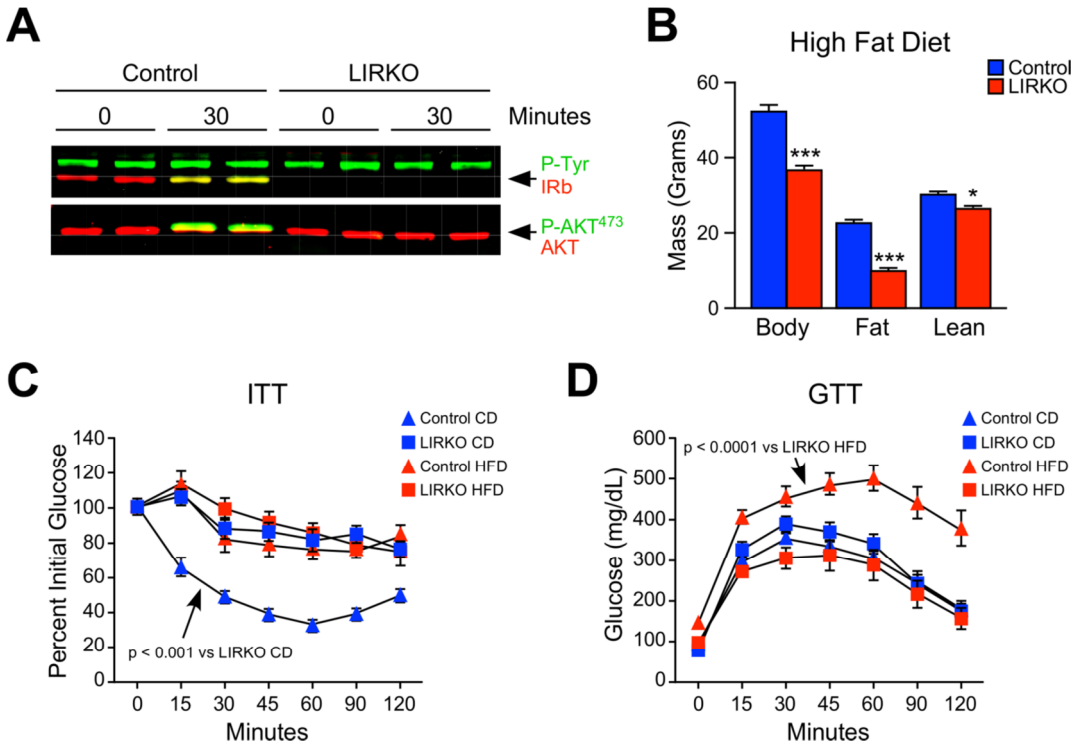


Figure 3-S4. Molecular and physiological analysis of LIRKO mice. (A) Control mice and mice lacking hepatic expression of Insr (LIRKO) were treated with insulin (5 U/kg) by intraperitoneal injection. Liver protein extracts were examined by immunoblot analysis by probing with antibodies to phospho-Tyrosine (P-Tyr), IRβ, pSer^{473}-AKT and AKT. (B) Control and LIRKO mice fed a HFD (16 weeks) were examined using ^{1}H-MRS to measure total body mass, fat mass, and lean mass (mean ± SEM; n = 10 (control) and 15 (LIRKO)). * $p < 0.005$; *** $p <$ 1e-07. (C, D) CD-fed and HFD-fed (16 weeks) control and LIRKO mice were examined at age 24 weeks using insulin tolerance tests (ITT, C) and glucose tolerance tests (GTT, D) (mean ± SEM; n = 7-14 (ITT) and 9-15 (GTT)). * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

CHAPTER 4

HYPER- AND HYPO- NUTRITION STUDIES OF THE HEPATIC TRANSCRIPTOME AND EPIGENOME SUGGEST THAT PPARα REGULATES ANAEROBIC GLYCOLYSIS

Diet plays a crucial role in shaping human health and disease. Diets promoting obesity and insulin resistance can lead to severe metabolic diseases, while calorie-restricted (CR) diets can improve health and extend lifespan. In this work, we fed mice either a chow diet (CD), a 16 week high-fat diet (HFD), or a CR diet to compare and contrast the effects of these diets on mouse liver biology. We collected transcriptomic and epigenomic datasets from these mice using RNA-Seq and DNase-Seq. We found that both CR and HFD induce extensive transcriptional changes, in some cases altering the same genes in the same direction. We used our epigenomic data to infer transcriptional regulatory proteins bound near these genes that likely influence their expression levels. In particular, we found evidence for critical roles played by PPARα and RXRα. We used ChIP-Seq to profile the binding locations for these factors in HFD and CR livers. We found extensive binding of PPARα near genes involved in glycolysis/gluconeogenesis and uncovered a role for this factor in regulating anaerobic glycolysis. Overall, we generated extensive transcriptional and epigenomic datasets from livers of mice fed these diets and uncovered new functions and gene targets for PPARα.

This chapter presents work that has been accepted as a manuscript at Scientific Reports.

Anthony R. Soltis*, Shmulik Motola[*], Santiago Vernia[*], Christopher W. Ng, Norman J. Kennedy, Simona Dalin, Bryan J. Matthews, Roger J. Davis, and Ernest Fraenkel
*Denotes equal contribution

Author contributions: S.M. performed RNA-Seq, DNase-Seq, and ChIP-Seq experiments, with assistance from S.D. and B.J.M. N.J.K. provided mice for sequencing experiments and analyzed body masses. A.R.S. analyzed RNA-Seq data, with assistance from S.M. A.R.S. and C.W.N. analyzed DNase-Seq data and performed motif analyses. A.R.S. and S.M. analyzed ChIP-Seq data. S.V. performed tolerance tests and *in vitro* and *in vivo* fenofibrate treatment experiments. A.R.S. and S.M. wrote the manuscript, with assistance from S.V. and N.J.K. E.F. and R.J.D. supervised the work.

Acknowledgments: The authors would like to thank Tali Mazor, Yoon Sing Yap, Candace Chouinard, members of the MIT BioMicro Center for their assistance with sequencing data collection. This work was supported in part by the Institute for Collaborative Biotechnologies through grant W911NF-09-0001 from the US Army Research Office (the content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred) and by NIH grants R24 DK-090963, R01 DK107220, and R01GM-089903 and sequencing support from P30-ES002109, as well as computing resources funded by the National Science Foundation under Award No. DB1-0821391.

Accession numbers: The raw and processed RNA-Seq, ChIP-Seq, and DNase-Seq data have been submitted to the NCBI Gene Expression Omnibus (GEO) as accession GSE47954.

4.1. INTRODUCTION

Diet plays a significant role in shaping human health and disease. Over nutrition leading to obesity can induce insulin resistance, a major human health concern that promotes the development of type 2 diabetes and some cancers [259-261]. In contrast, caloric restriction can

extend lifespan, improve insulin sensitivity, and delay the onset of age-related diseases, such as diabetes, cardiovascular disease, and neoplasia [262, 263]. While the broad contrasts between high-fat diet feeding and calorie restriction are well established, the underlying molecular processes that drive these physiological and metabolic differences are incompletely understood.

The liver is a critical regulator of metabolism and is sensitive to dietary changes. The liver maintains normal glucose homeostasis by suppressing hepatic gluconeogenesis in response to insulin following feeding, while promoting glucose production during fasting [15, 41]. High-fat diet induced obesity and insulin resistance disrupts these hepatic mechanisms and promotes hyperglycemia [30]. Caloric restriction, however, lowers liver fat accumulation and improves hepatic glucose regulation in obese humans [264, 265] and reduces the expression of stress and inflammatory genes in mouse livers, which may contribute to the anti-aging effects associated with this diet [266]. The liver, therefore, is a critical driver of the body's response to dietary challenges. Thus, analysis of hepatic responses to dietary extremes may enhance our understanding of how diet shapes overall human health.

In this study, we profiled transcriptional and epigenomic landscapes in the livers of mice fed either a standard laboratory chow diet (CD), a long-term (16 week) high-fat diet (HFD) to induce obesity and insulin resistance, or a nutrition-restricted diet to model caloric restriction (CR). Overall, we present a comprehensive analysis of diet-induced effects on mRNA expression and chromatin accessibility in the mouse liver following HFD and CR. We find that calorie restriction and high fat feeding have common and independent epigenetic and transcriptomic signatures. We also show that PPARα activation underlies both extreme metabolic situations and identify new PPARα targets that regulate glucose metabolism.

4.2. RESULTS

4.2.1. High-fat diet and calorie restriction induce extensive changes in hepatic gene expression

We examined mice following a long-term (16 week) high-fat diet (HFD) or a calorie restricted (CR) feeding protocol. As anticipated, mice fed a HFD gained body mass while CR mice lost mass compared to chow diet (CD)-fed controls (p < 5e-5, two-sided t-tests) (**Figure 4-1A**). We assessed glucose homeostasis in HFD mice compared to controls using tolerance tests for glucose (GTT, **Figure 4-1B**), insulin (ITT, **Figure 4-1C**), and pyruvate (PTT, **Figure 4-1D**) and confirmed that mice fed a HFD are strongly insulin resistant glucose intolerant.

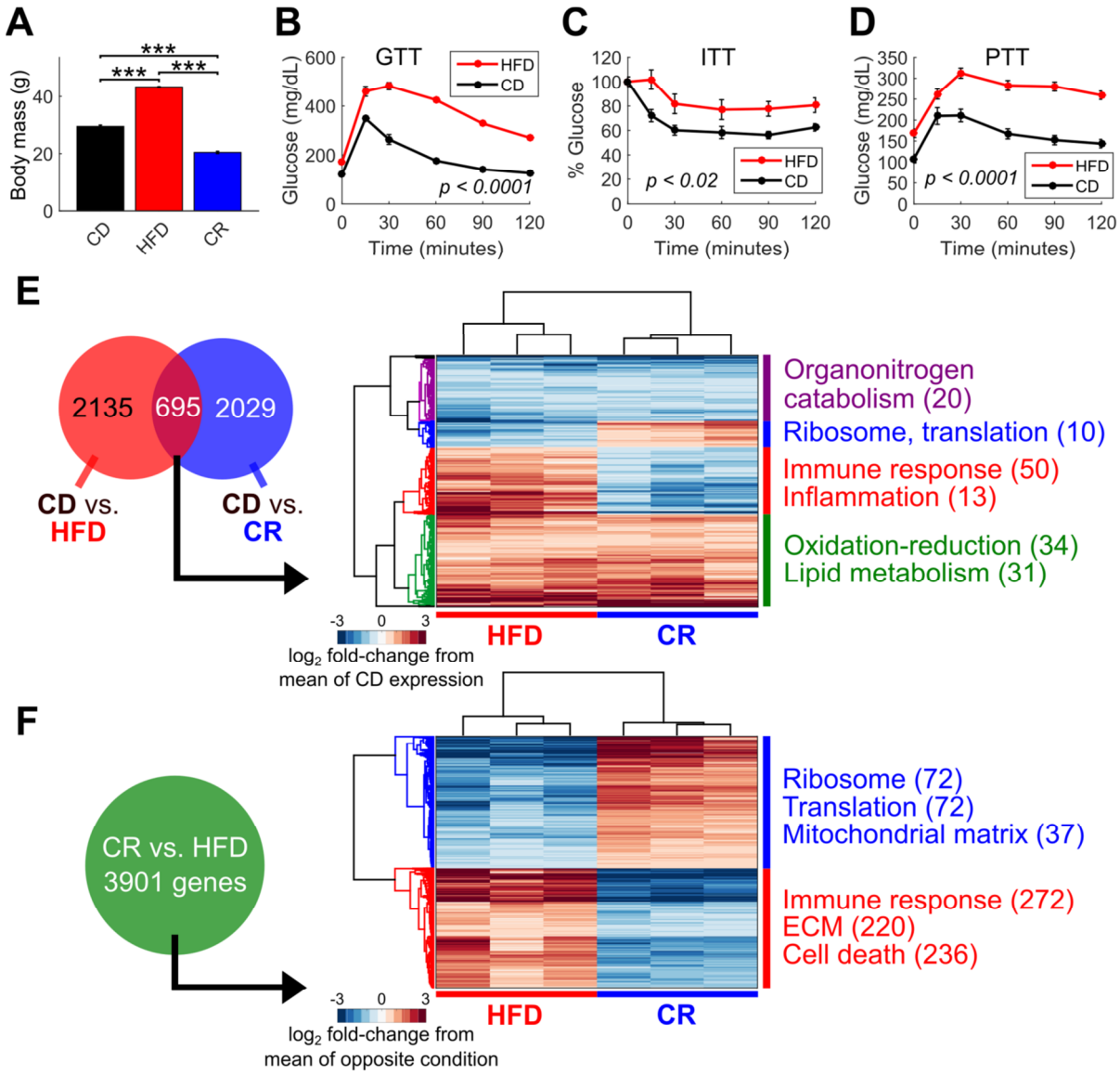


Figure 4-1. High-fat diet and calorie restriction alter body mass and induce extensive hepatic transcriptional changes.

Figure 4-1 (continued). High-fat diet and calorie restriction alter body mass and induce extensive hepatic transcriptional changes. (A) HFD and CR feeding increase and decrease overall mouse body mass, respectively, compared to CD (n = 12, 10, and 12 for CD, HFD, and CR, *** $p <$ 5e-5, two-sided t-tests). (B-D) HFD induces insulin resistance and alters glycemic regulation as assessed by (B) glucose tolerance test (GTT), (C) insulin tolerance test (ITT), and (D) pyruvate tolerance test (PTT) (*p*-values from t-tests of area under the curve measurements, n = 30, 25, and 23 for CD and n = 37, 27, and 23 for HFD). (E) Venn diagrams show numbers of genes differentially expressed between CD and HFD livers (red circle) as well as CD and CR livers (blue circle). The overlap region shows 695 genes that are differentially expressed in both CR and HFD compared to CD. The clustergram shows these 695 overlapping genes that are up-regulated by both HFD and CR (255 genes), down-regulated by both CR and HFD (183 genes), up-regulated in HFD and down-regulated by CR (186 genes), and up-regulated in CR but down-regulated in HFD (71 genes), along with gene ontology and pathway enrichment terms. The numbers reflect the numbers of genes in each group that are annotated to each term. Values are $\log_2$ fold-changes for individual replicate expression levels (in FPKM) versus the mean CD expression level. (F) 3,901 genes are differentially expressed between CR and HFD livers (green circle). The clustergram shows individual replicate gene expression levels as $\log_2$ fold-change compared to the mean expression level for the opposite condition (CR or HFD). The numbers reflect the numbers of genes in each group that are annotated to each term.

We comprehensively quantified the hepatic transcriptomic landscapes of these mice using RNA-Seq (**Figure 4-S1B**). Both HFD and CR induced widespread changes in hepatic gene expression compared to CD, with 2,830 and 2,724 genes differentially expressed by the two conditions, respectively (FDR < 0.05, absolute $\log_2$ fold-change > 0.5) (**Figure 4-1E**). HFD induced the expression of genes involved in immune response (FDR < 6.4e-22, e.g. *Ccr1*, *Ccr2*, *Cd36*, *Tlr1*), lipid metabolism (FDR < 8e-6, e.g. *Abcd1*, *Apoa4*, *Cyp17a1*, *Srebf1*, *Thrsp*), stress response (FDR < 1.3e-5, e.g. *Anxa1*, *Axl*, *Car3*, *Hif1a*, *Jak2*), and cell death (FDR < 6e-4, e.g. *Bak1*, *Casp7*, *Jun*), among others. CR up-regulated genes are involved in cholesterol metabolism (FDR < 2.5e-11, e.g. *Cebpa*, *Dhcr7*, *Hmgcr*, *Ldlr*) and mitochondria (FDR < 7e-7, e.g. *Atp5e*, *Cox5a*, *Mrps24*), among other processes.

We found a significant set of 695 genes ($p <$ 3.6e-14, hypergeometric test of 695 overlapping genes) that are differentially regulated by both HFD and CR compared to CD, including 255

genes up-regulated by both HFD and CR, 183 down-regulated by both, 186 up-regulated in HFD and down-regulated by CR, and 71 up-regulated in CR but down-regulated in HFD (**Figure 4-1E**). Of note, the majority of these genes (438 or ~63%) change in the same direction compared to CD (p < 2e-14, Fisher's exact test). The first set of genes (up-regulated in both conditions) is enriched in processes related to oxidation-reduction (FDR < 0.004) and lipid metabolism (FDR < 0.021). This latter category includes genes involved in cellular fatty acid synthesis (e.g. *Fads1*, *Fads2, Fasn*, *Scd1*), lipid and cholesterol production (*Dhcr24*, *Nsdhl*, *Srebf1*), triglyceride synthesis (*Thrsp*), and peroxisomal import of free fatty acids (*Abcd1*, *Abcd2*). We note that expression changes in oxidation-reduction and lipid metabolism in CR mice are not a consequence of any increases in consumed dietary fat content, as the CR diet contains a similar percentage fat content to the CD and because CR mice consumed overall less food, and therefore less fat, than both the CD and HFD mice. The second set of overlapping genes (down-regulated in both conditions) is enriched in organonitrogen catabolism (FDR < 0.02, e.g. *Aass*, *Agxt*, *Cbs*, *Kynu*, *Pnp*). Genes up-regulated by HFD but down-regulated by CR are involved in immune response (FDR < 6.9e-8, e.g. *Apoa4*, *C1qa,b,c*, *Gas6*) and inflammation (FDR < 6.5e-3, e.g. *Aif1*, *Axl*, *Csf1*, *Tgfb1*), while genes up-regulated by CR but down-regulated by HFD compared to CD are involved in translation and ribosomal composition (FDR < 0.019, e.g. *Rpl37*, *Rps15a*, *Rps28*, *Rps3*). This analysis highlights a common set of genes altered by both conditions that, in a majority of cases, are altered in the same way, a surprising result given the differences in the overall metabolic states of CR and HFD mice.

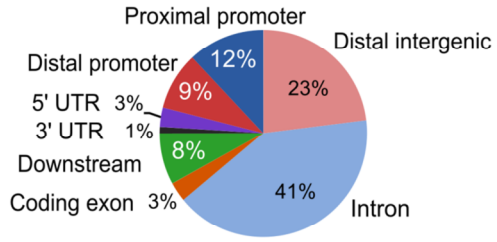
We next compared the CR and HFD liver RNA-Seq samples to directly contrast the two dietary extremes. We found in total 3,901 differentially expressed genes, with 1,857 genes up-regulated in HFD and 2,044 up-regulated by CR (**Figure 4-1F**; **Figure 4-S1A** for qPCR validation of select genes). Similar to our other comparisons of the gene sets altered by these diets, genes up-regulated by CR are enriched in processes related to ribosomes, mitochondria, translation, and tRNA processing, while HFD-induced genes are enriched in immune responses, extracellular matrix components, and cell death. Thus, although we found evidence for genes regulated similarly following CR and HFD (**Figure 4-1E**), in general these two dietary extremes induce distinct gene expression programs.

4.2.2. DNase-Seq and motif analyses identify PPARα and RXRα as common regulators of HFD and CR-induced gene expression in liver

Given the widespread hepatic transcriptional changes induced by both HFD and CR feeding, we performed DNase-Seq on the livers of CD, HFD, and CR mice in order to uncover accessible regulatory regions throughout these genomes that likely harbor regulatory proteins associated with the transcription of these differential genes. Globally, we found high correlations (r = 0.76 – 0.84) between hypersensitive regions identified in the livers of the mice on the three diets (**Figure 4-S2A-C**). For subsequent analyses, we merged the regions identified in all three diets into a set of 92,626 total hypersensitive sites to maximize the search space for regulators (**Figure 4-2A**). We mapped each of these regions to known gene coordinates within the mouse genome and found that the majority of these regions reside within introns (41%). Additional near-gene sites include: proximal promoters (12%), distal promoters (9%), sites downstream of gene bodies (8%), coding exons (3%), 5' UTRs (3%), and 3' UTRs (1%). The remaining sites map to distal intergenic regions linearly distant from known gene boundaries (23%). Thus, the majority (77%) of identified hypersensitive regions appear in or near annotated gene boundaries throughout the mouse genome.

As specific examples, we found hypersensitive regions across the conditions near the gene *Cyp2b10*, which is a known target of the nuclear hormone receptors CAR and RXR [267, 268] (**Figure 4-S2D**). Additionally, we found a number of hypersensitive sites near and within the introns of the gene *Abca1*, which is a known target of RXR and LXR in the liver [269] (**Figure 4-S2E**). We performed direct motif analysis on the hypersensitive regions near these select genes and indeed found enrichment for the RXR:LXR motif (**Figure 4-S2F**). Thus, the use of motif analysis on the hypersensitive regions near genes altered by diet could reveal regulators associated with these changes.

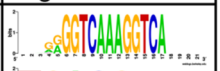
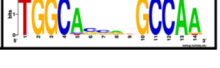
A 92626 Hypersensitive sites



B

Gene set	# hypersensitive sites +/- 10 kb TSS *low* CpG	# hypersensitive sites +/- 10 kb TSS *high* CpG
[1]CD vs. HFD	5881	3930
[2]CD vs. CR	6405	3969
Overlap 1 & 2	1832	934
CR vs. HFD	7788	5289

C - Top motifs in *low* CpG regions near genes

Motif	Factors	Logo
Nuclear hormone receptors, DR-1	*Ppara, Pparg, Rxra, Hnf4a*	
ATF/CREB	*Atf1-7, Creb1*	
HNF1 dimers	*Hnf1a, Hnf1b*	
FXR:RXR inverted repeat	*Nr1h4, Rxra*	
Fork head	*Foxa, Foxo, Foxj, Foxk*	
Nuclear factor 1 dimers	*Nfia, Nfib, Nfic, Nfix*	

D - Top motifs in *high* CpG regions near genes

Motif	Factors	Logo
ATF/HLF	*Atf2, Hlf, Nfil3*	
Dimeric nuclear respiratory factor	*Nrf1*	
ELK/ETS/ETV	*Elk, Ets, Etv*	
E2F, DP-1	*E2f1-8, Tfdp1-2*	
AHR/EGR	*Ahr, Egr1-4*	
Heterotrimeric trans. factors	*Nfya, Nfyb, Nfyc*	

Figure 4-2. DNase-Seq reveals regulatory regions across liver genomes and motif analyses identify potential transcriptional regulators. (**A**) We found 92,626 hypersensitive regions among CD, HFD, and CR livers. The majority of identified sites reside within annotated gene introns (41%) as well as other near-gene locations. Regions were mapped near genes according to: proximal promoters – within 200 bp of gene TSS; distal promoters – within 5 kb upstream of gene; downstream – within 5 kb downstream of gene end; introns, exons, 5' UTR, and 3' UTR – if region intersected one of these features; and distal intergenic – outside 5 kb window around gene. (**B**) The numbers of identified hypersensitive regions near differential gene sets (first column) within +/- 10 kb of gene transcription start sites (TSS) in low (middle column) and high (right column) CpG content sequence sets are presented. (**C-D**) The most enriched DNA binding motifs near all gene sets are shown for low (**C**) and high (**D**) CpG content sequences.

We examined the discovered DNase hypersensitive regions near the transcription start sites (TSSs, +/- 10 kb for this analysis) of our differential gene sets for enriched transcriptional regulator motifs (**Figure 4-2B**). We divided the hypersensitive sites near these gene sets into low (< 0.5) and high (> 0.5) CpG content sequences and assessed motif enrichment in both sets of sequences (see Methods). We found distinct motif enrichments in low versus high CpG content hypersensitive sequences, but consistently observed nearly the same motif enrichments (and rankings) across all the gene sets for the two sets of sequences. In low CpG content regions, we observed strong enrichments near all the gene sets for nuclear hormone direct repeat 1 motifs,

corresponding to the factors PPARα, PPARγ, RXRα, and HNF4α, as well as ATF/CREB, HNF1 dimer (HNF1α and HNF1β), FXR:RXR inverted repeat, fork head factor (FOXA, FOXO, FOXJ, FOXK, etc. factors), and nuclear factor 1 dimer (NFIA, NFIB, NFIC, and NFIX) motifs (**Figure 4-2C**). In high CpG content regions, we also observed motif enrichment for ATF factors, though with more preference for thymine as opposed to guanine in the second position of the motif (ATF, HLF factors), as well as dimeric nuclear respiratory factor (NRF1), ELK/ETS/ETV, E2F (E2F and DP-1 factors), AHR/EGR, and heterotrimeric transcription factor (NFYA, NFYB, NFYC) motifs (**Figure 4-2D**). We only found modest enrichments for factors when comparing conditions against one another, e.g. low CpG content regions near up-regulated genes in CR versus regions near genes up-regulated in HFD. This observation is likely due to the fact that we saw such strong enrichments for the same factors in low and high CpG content regions regardless of the gene sets tested. Thus, these factors likely play multiple roles in different contexts to regulate the gene expression programs we observed across the various diets. Given the strong enrichment for nuclear hormone receptor motifs in the low CpG content regions we analyzed, we chose to investigate further the genome-wide binding profiles for the factors PPARα and RXRα to examine their roles in regulating CR and HFD hepatic gene expression.

4.2.3. ChIP-Seq profiling of PPARα and RXRα binding in CR and HFD livers reveals extensive genome-wide regulation and uncovers novel targets

Our motif analysis strongly suggested that PPARα and RXRα, two transcription factors prominently expressed in liver [270, 271], contribute to the differential expression of genes in the livers of mice fed either a high fat or calorie restricted diet. We also found significant enrichment for a set of 228 known PPARα target genes among all the differential genes (hypergeometric p-values < 5e-4) [272]. For example, 27 of the 695 genes differential in both CR and HFD livers compared to CD (**Figure 4-1E**) are among this set of known PPARα targets (p < 3.7e-7). We thus used ChIP-Seq with specific antibodies against these factors (**Figure 4-S3A**) to profile their genome-wide binding profiles in CR and HFD livers.

As anticipated from our motif analyses, our ChIP-Seq datasets confirmed that both PPARα and RXRα bind extensively near genes in these livers (**Figure 4-S3B**). Overall, we detected more

RXRα binding than PPARα, likely due to the lower obtained sequencing depth from PPARα samples. Over all binding sites for each factor, we detected some form of the PPAR:RXR heterodimer motif (direct repeat 1) in 91% and 90% of PPARα and RXRα regions, respectively; thus, the majority of identified binding sites contain an expected motif for these factors, though ~10% of these sites likely reflect alternative binding mechanisms (e.g. via other DNA-binding co-regulatory proteins). PPARα binding sites mapped to 1,253 and 2,320 annotated genes in CR and HFD, respectively, while RXRα enriched regions mapped 3,381 and 4,767 genes (+/- 10 kb window). The genome-wide binding distributions for these factors also closely mirror those observed in our DNase-Seq experiments, with the majority of binding regions located in introns (42-44%) as well as other near-gene regions (**Figure 4-S3B**, left and middle columns). 23-32% of all binding sites were classified as distal intergenic. We also searched for regions in which we found proximal binding events for both factors (peak summits within +/- 100 bp) and found 2,831 and 8,838 such regions in CR and HFD livers. The genome-wide binding locations for these regions were similar to those observed for the individual factors (**Figure 4-S3B**, right column).

We used the uncovered binding events to identify known and novel genes that are likely regulated by these factors. PPARα, typically bound as a heterodimer with RXRα, is a well-characterized regulator of lipid metabolism [273], and we saw strong enrichment for such metabolic processes in up-regulated genes in both CR and HFD livers (**Figure 4-1E**). Consistent with this, we identified binding events near the transcription start sites of genes involved in various lipid metabolic processes which are known to be regulated by PPARα/RXRα [272], including *Acadl* (involved in mitochondrial β-oxidation), *Cpt2* (involved in mitochondrial oxidation of long-chain fatty acids), *Fabp1* (involved in fatty acid uptake and transport), and *Fgf21* (involved in fatty acid oxidation and ketogenesis) (**Figure 4-3A**). Among these, we found binding evidence for both PPARα and RXRα near *Fgf21* in HFD only (**Figure 4-3A**, bottom right). This result is consistent with our RNA-Seq data in that *Fgf21* is up-regulated in HFD livers compared to CR ($\log_2$ fold-change of 2.9, FDR < 4.5e-7).

Our analyses identified several novel targets of PPARα and RXRα, including *Crtc2* and *Nfic* (**Figure 4-3B**). *Crtc2* is a known co-regulator of glucose metabolism [274]. We identified

binding events for both factors across the two diets at the promoter of this gene. We also highlight binding near *Nfic*, a gene also up-regulated in HFD livers compared to CR, which has up-stream binding events for PPARα in HFD only, as well as clear binding peaks for RXRα alone at its TSS in both CR and HFD. Thus, our profiling of PPARα and RXRα in CR and HFD-

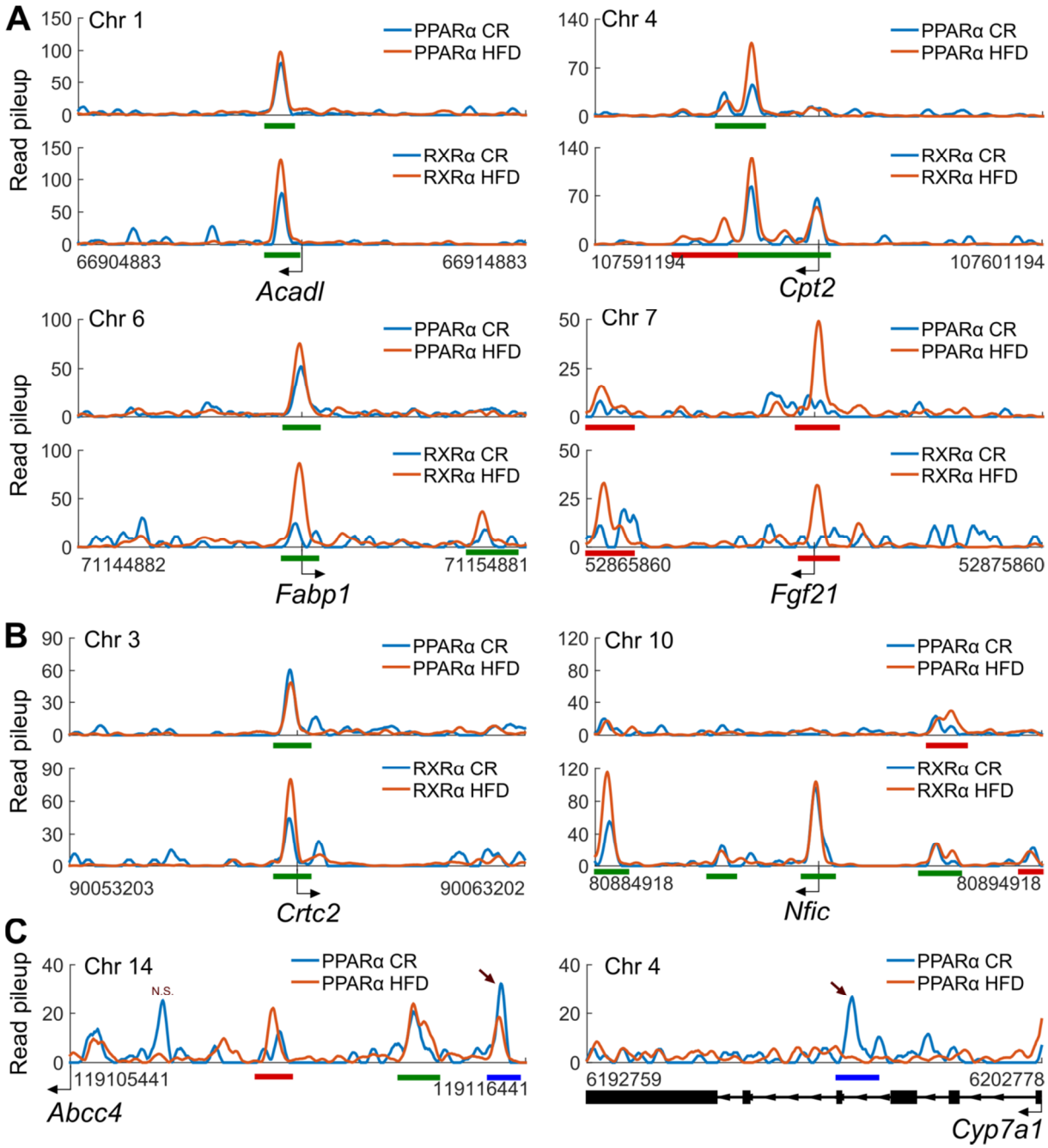


Figure 4-3. ChIP-Seq of PPARα and RXRα transcription factors in CR and HFD livers reveals extensive binding near known and novel regulated genes.

Figure 4-3 (continued). ChIP-Seq of PPARα and RXRα transcription factors in CR and HFD livers reveals extensive binding near known and novel regulated genes. (A) The binding profiles (+/- 5 kb gene TSS) for known PPARα and RXRα targets *Acadl*, *Cpt2*, *Fabp1*, and *Fgf21* in CR and HFD livers are shown. (B) The binding profiles (+/- 5 kb gene TSS) for novel PPARα and RXRα targets *Crtc2* and *Nfic* in CR and HFD livers are shown. (C) The binding profiles for PPARα near the differentially expressed genes *Abcc4* and *Cyp7a1* (CR vs. HFD) that contain differential binding events between the same two diets in our ChIP-Seq data. Arrows indicate differential binding regions; N.S. stands for not significant. Read pileup refers to extended, normalized, and smoothed read pileup counts extracted from concatenated pools of aligned reads for the biological replicates for each factor (see Methods). Green lines indicate significantly called peaks in both CR and HFD. Red and blue lines indicate significantly called peaks in HFD and CR, respectively.

fed mouse livers revealed binding events near many genes known to be regulated by these factors, while also uncovering new genes not previously characterized as targets of these factors.

Finally, we tested our PPARα and RXRα ChIP-Seq datasets for evidence of differential binding between CR and HFD livers. We observed a small set of statistically significant differential binding events between the diets for RXRα regions (381 regions, 1.2% of total), even though we identified roughly two times as many called RXRα peaks in HFD compared to CR (**Figure 4-S3B**). This result is likely due to thresholding differences during binary peak calling (e.g. due to sequencing depth) which do not always manifest as true statistical differences when comparing read counts in these regions directly. 112 of these 381 differential peaks mapped within +/- 20 kb of 103 differential genes between CR and HFD livers. We saw more evidence for differential binding of PPARα between CR and HFD, with 1,201 (9.3% of total) identified peaks showing significant differential enrichment. Only 307 of these, however, mapped to a gene differentially expressed between CR and HFD, covering 284 of the nearly 4,000 potential differential genes. Among these, we observed a differential peak ~10 kb upstream of the *Abcc4* gene promoter that shows lower enrichment in HFD compared to CR (**Figure 4-3C, left**). Indeed, *Abcc4* is expressed significantly lower (~ -1 $\log_2$ fold-change) in HFD compared to CR in our RNA-Seq data. As another example, we found a differential peak with higher enrichment in CR within the gene body of *Cyp7a1*, which is also expressed higher in CR compared to HFD by RNA-Seq (**Figure 4-3C, right**). Though we did not detect many *differential* binding events near these genes, we did detect many binding events in general for these factors near a substantial number

of the differential genes. 4,060 PPARα sites map to 1,879 of these genes and 10,271 RXRα peaks map to 2,994. Thus, we found specific instances of differential PPARα and RXRα binding near differential genes between CR and HFD, though such differences only explain small fractions of the total differential gene pools. These results suggest that these factors, given that they indeed bind near many of the differential genes even if the degrees of binding do not measurably change, regulate gene expression differences by mechanisms other than differential binding (e.g. due to differential activity levels or co-factor/co-repressor binding events), though some genes may be more sensitive to differential binding events by other factors.

4.2.4. mRNA expression, binding data, and fenofibrate-treated primary hepatocytes further suggest a role for PPARα in regulating glucose metabolism

While PPARα has extensively been characterized as a regulator of lipid metabolism, there is evidence that this transcription factor plays a role in regulating glucose metabolism [272, 275-277]. In particular, PPARα knock-out mice show severe hypoglycemia and depleted hepatic glycogen stores during fasting [278]. Moreover, PPARα has been shown to regulate the gluconeogenic genes *G6pc*, *Pck1*, and *Pcx*, the glycerol converting genes *Gpd1* and *Gpd2*, and the pyruvate dehydrogenase inhibitor *Pdk4* [272, 277, 279]. Indeed, we detected PPARα binding events near the transcription start sites or within the bodies of these genes.

We examined genes in the canonical gluconeogenesis/glycolysis pathway for evidence of PPARα binding and found events near nine genes (of fourteen queried) encoding enzymes in this pathway (**Figure 4-4A**). Interestingly, we found that *Aldob*, *Fbp1*, *Fbp2*, *Pck1*, and *Pklr* not only bind PPARα, but are sensitive to PPARα activation [275] (**Figure 4-4A**, blue highlighted genes). Furthermore, our RNA-Seq data demonstrate PPARα-bound genes are regulated by feeding a HFD or CR, including *Gck* and *Pklr* that are up-regulated by CR and HFD, *G6pc* and *Gapdh* that are down-regulated by CR, and *Eno3* that is down-regulated in HFD (**Figure 4-4A**, colored bars). Thus, PPARα likely influences diet-induced expression changes in these genes.

To further test the role of PPARα in regulating hepatic glucose metabolism, we treated mouse primary hepatocytes with fenofibrate, a PPARα agonist, and measured glycolytic rates. PPARα

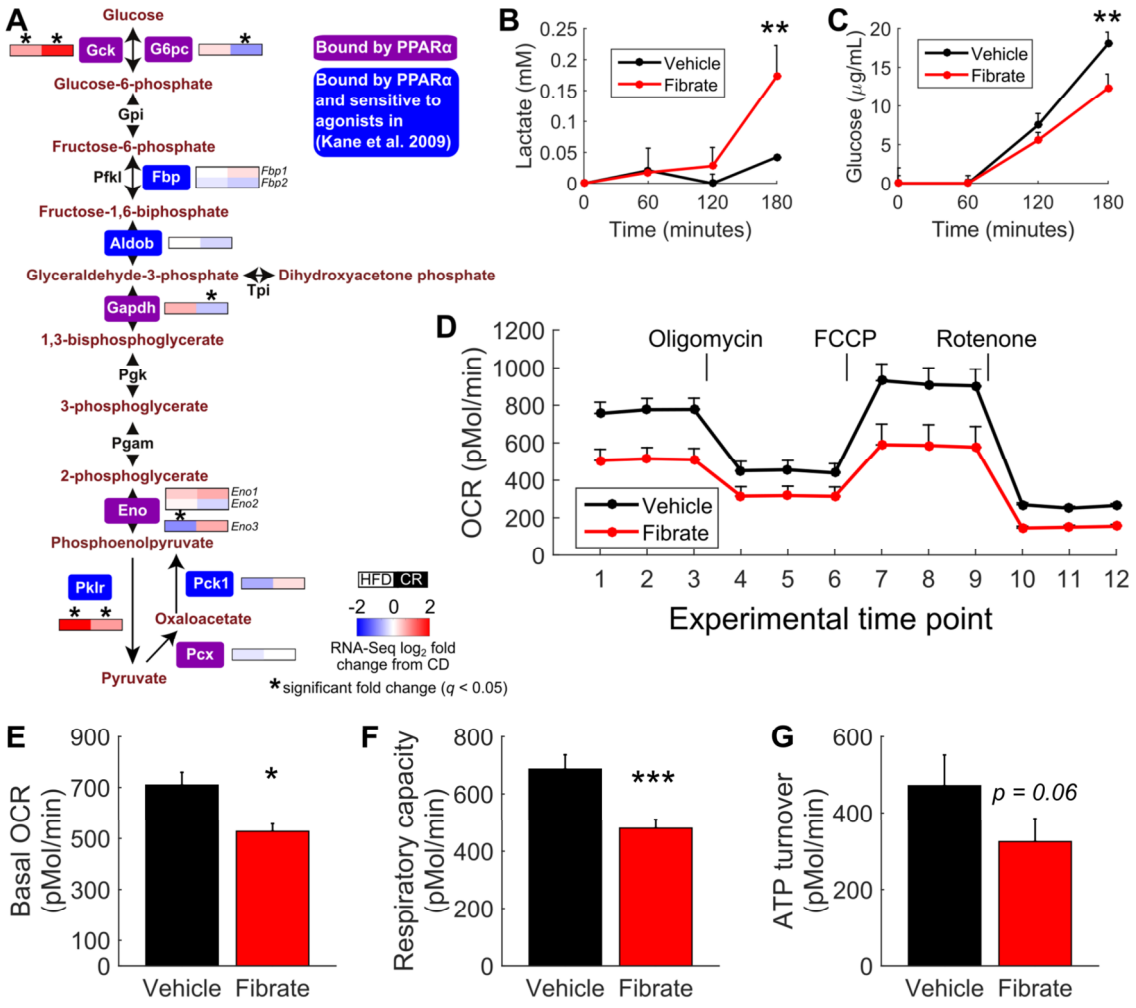


Figure 4-4. PPARα binds extensively near genes involved in gluconeogenesis/glycolysis in CR and HFD livers and activation by fenofibrate enhances anaerobic glycolysis in primary hepatocytes. (A) Canonical gluconeogenesis/glycolysis pathway highlighting genes bound by PPARα (purple outline boxes) and genes both bound by PPARα in our dataset and sensitive to fibrate in Kane et al. (2009) study (blue outline boxes). The two-element color bars near the bound genes represent the $\log_2$ fold-changes in mRNA expression (by RNA-Seq) for these genes in HFD and CR livers, respectively, versus CD. * indicates statistically significant changes ($q < 0.05$). (B) Lactate production in mouse primary hepatocytes following vehicle (black line) or fenofibrate (red line) treatment in the presence of glucose. (C) Glucose production in the presence of lactate/pyruvate as a gluconeogenic source following vehicle (black line) or fenofibrate (red line) treatment. (D) Oxygen consumption rate (OCR) assessed in the presence of glucose following vehicle (black line) or fenofibrate (red line) treatment. OCR also assessed following oligomycin, FCCP, and rotenone drug treatments. (E-G) Assessment of basal OCR (E), respiratory capacity (F), and ATP turnover (G) in primary hepatocytes following vehicle or fenofibrate treatment.

activation in hepatocytes cultured with glucose as a fuel displayed a significant increase in lactate production, suggestive of an increase in glycolytic flow (**Figure 4-4B**). Consistent with this result, we observed decreased glucose production in the presences of lactate/pyruvate as a gluconeogenic source in fibrate-treated hepatocytes (**Figure 4-4C**). These results suggest that PPARα enhances glycolysis, leading to non-oxidative conversion of glucose to lactate. To test this hypothesis, we assessed the oxygen consumption rate (OCR) in primary hepatocytes using glucose as a fuel (**Figure 4-4D**). OCR was consistently lower in fenofibrate-treated hepatocytes, even in the presence of oxygen consumption inhibitors (oligomycin and rotenone) and enhancers (FCCP). We observed reduced basal OCR (**Figure 4-4E**) and maximal respiratory capacity (**Figure 4-4F**), as well as lower ATP turnover (**Figure 4-4G**), in fenofibrate-treated primary hepatocytes compared to vehicle controls, confirming that PPARα activation decreases oxidative metabolism of glucose. These results, together with our binding data and RNA-Seq results in CR and HFD livers, further stress a role for PPARα in regulating glucose metabolism.

4.2.5. *In vivo* fenofibrate treatment confirms role of PPARα regulation near genes involved in glucose metabolism in liver

We next tested the effect of *in vivo* fenofibrate treatment on specific PPARα targets identified by our ChIP-Seq data. We treated mice for two weeks with either vehicle or fenofibrate and measured hepatic gene expression of PPARα targets using quantitative PCR. We found significant up-regulation of well-characterized PPARα target genes following fibrate treatment, including *Acox1*, *Ehhadh*, and *Pdk4* (**Figure 4-5A**). We next tested several novel PPARα targets identified from our ChIP-Seq data analysis. In keeping with our identification of a role for PPARα in regulating glucose metabolism, we found binding sites near the genes *Fbp1* and *Gck* in both CR and HFD and *Pklr* in HFD. Following *in vivo* fenofibrate treatment, the expression levels of these genes were significantly repressed, providing further support that these are regulated targets of PPARα (**Figure 4-5B**). We also tested other novel targets bound in our ChIP-Seq data, including *Aldh1*, *Aldh2*, *Eno1, Pcx*, and *Sirt3*; however, these were not significantly altered following fibrate treatment by qPCR, suggesting that additional mechanisms are necessary to control their expression *in vivo*.

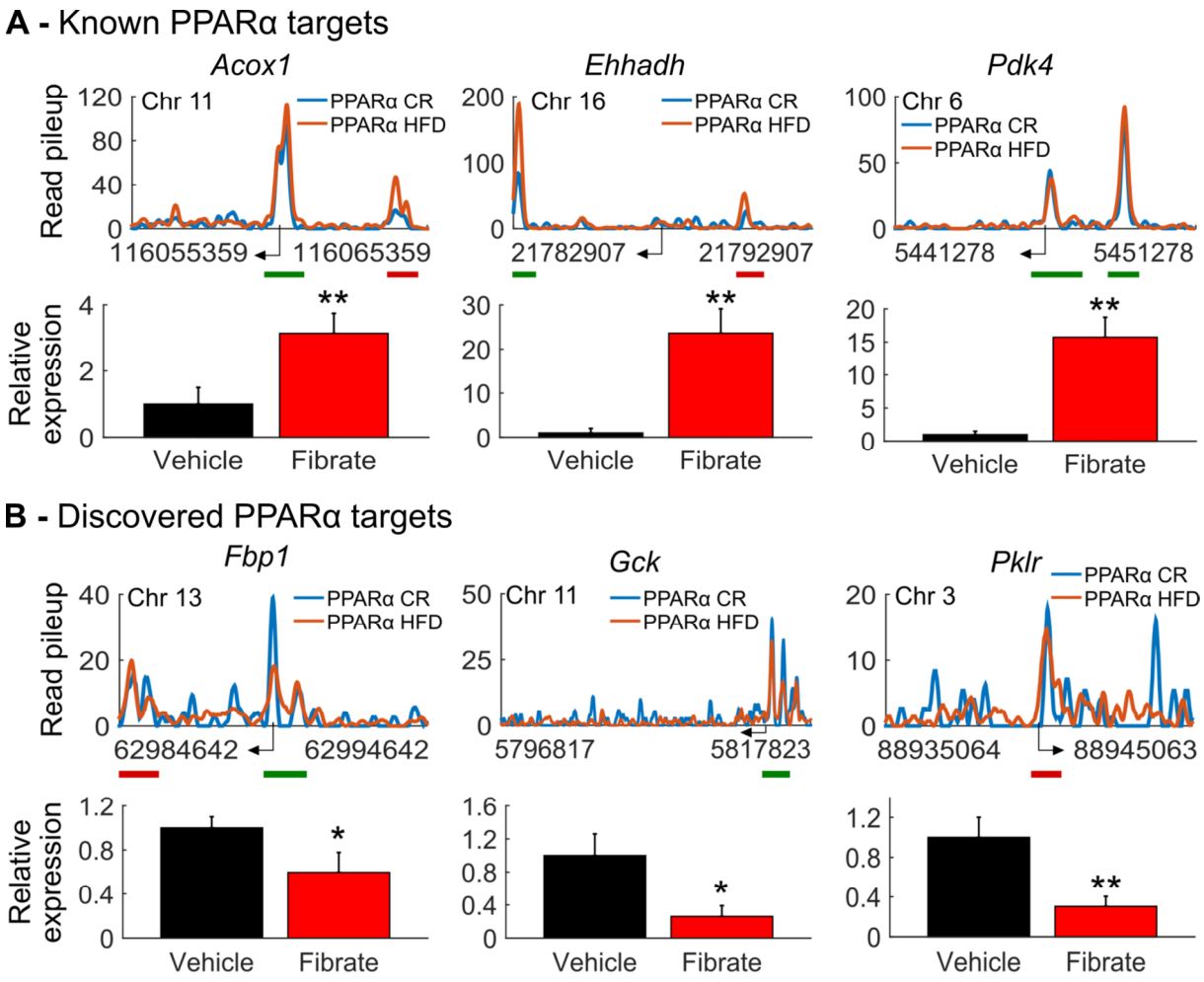


Figure 4-5. Identified targets are regulated in vivo by PPARα. (**A**) Read pileup profiles for CR and HFD PPARα ChIP-Seq near known regulated genes with corresponding *in vivo* qPCR results following fenofibrate treatment. (**B**) Read pileup profiles for CR and HFD PPARα ChIP-Seq near novel regulated genes with corresponding *in vivo* qPCR results following fenofibrate treatment. Read pileup refers to extended, normalized, and smoothed read pileup counts extracted from concatenated pools of aligned reads for the biological replicates for each factor (see Methods). Green lines indicate significantly called peaks in both CR and HFD. Red and blue lines indicate significantly called peaks in HFD and CR, respectively. * $p < 0.05$, ** $p < 0.01$.

4.3. DISCUSSION

In this study, we examined the hepatic transcriptional and epigenetic landscapes of mice fed chow, high-fat, and calorie-restricted diets. Joint analysis of our epigenetic data with regulatory protein DNA binding motif data revealed a common set of transcription factors that may regulate

the genes altered by these diets. In particular, we found strong enrichments for the PPARα and RXRα motifs near all the identified differential gene sets. We further investigated these findings with direct ChIP-Seq experiments for these factors and found that they do indeed bind extensively near these genes throughout the genomes of HFD and CR mice, further suggesting extensive roles for these factors in the hepatic response to dietary challenges. We particularly focused on the role of PPARα in regulating glucose metabolism in the liver and found extensive binding near genes encoding proteins involved in metabolism of carbohydrates. We validated these findings by treating primary mouse hepatocytes with fenofibrate to stimulate PPARα activation, discovering that activation of this factor enhances anaerobic glycolysis. We also performed *in vivo* fenofibrate treatment experiments in mice and, using quantitative PCR, validated several novel gene targets for PPARα involved in glucose handling. Overall, we present a comprehensive analysis of the effects of high-fat feeding and caloric restriction on mouse hepatic transcriptomics and epigenomics, along with new insights into how the divergent physiological and metabolic states induced by these diets are regulated at the level of transcription.

Our transcriptional profiling data revealed extensive changes in gene expression induced by both HFD and CR compared to CD, as well as many changes between HFD and CR directly. Interestingly, we observed a significant set of 695 genes that change in both extreme diets compared to controls, with 438 of these changing in the same direction. Genes that are up-regulated in both HFD and CR are enriched in oxidation-reduction and lipid metabolic processes, while genes down-regulated in both conditions are enriched in organonitrogen catabolism. Our data thus suggests that some processes and pathways, e.g. fatty acid synthesis, are commonly utilized by the liver in response to divergent dietary challenges. To this end, HFD induces unsaturated fatty acid and triglyceride synthesis to accumulate fat [280] while CR induces adipose and liver enzymes involved in fatty acid metabolism, including *Fasn* and *Srebf1* which were significantly up-regulated by both diets in our data, to reduce oxidative stress and to induce energy production via β-oxidation [281]. Thus, the liver can co-opt specific pathways for purposes suitable to the needs of various external challenges, in this case dietary changes.

The majority of transcriptional changes induced by HFD and CR, however, are divergent. Specifically up-regulated at the mRNA level in HFD livers are genes involved in immune responses, inflammation, extracellular matrix, and cell death, consistent with anticipated complications resulting from obesity-induced insulin resistance [56, 282, 283]. CR livers induced genes related to ribosomes, translational processes, and mitochondria. The inflammatory state observed in HFD livers likely maintains complications related to the insulin resistant state, whereas the reduction in genes associated with these processes in CR, whether compared to CD or HFD, may contribute to the beneficial effects of caloric restriction.

Epigenetic data from DNase-Seq experiments allowed us to map the landscapes of accessible regulatory regions throughout the livers of mice fed each of these diets. We found that most of these accessible regulatory regions reside either within or proximal to known gene boundaries (77%), with the majority residing in introns (41%). We used sequence analysis of these accessible regulatory regions with known DNA-binding motif preferences for regulators to infer factors that are likely associated with transcription of the genes altered by diet. We separated sequences into low and high CpG content sets and looked for motif enrichments across several differential gene sets. While we did find factor enrichment differences between the low and high CpG content sets of sequences, as can be anticipated, we did not find many differences in motif enrichments between the various gene sets. These results suggest that common sets of regulatory proteins are utilized for numerous purposes in the liver. We identified strong enrichments for nuclear hormone receptors, ATF/CREB, and HNF1 factors in low CpG content regions, whereas we found nuclear respiratory factor and ELK/ETS/ETV factor enrichments in high CpG content regions (among others). The strong enrichment of nuclear hormone receptor factors led us to examine the binding profiles for some of these factors more specifically in HFD and CR livers.

We profiled PPARα and RXRα binding throughout the livers of HFD and CR mice using ChIP-Seq. Overall, we found extensive binding for these factors across the genomes as suggested by our motif analyses. We confirmed many known binding sites for these factors near the transcription start sites of specific genes, but also found several novel binding events near genes not known to be regulated by PPARα or RXRα (e.g. *Crtc2* and *Nfic*). We also directly compared binding events for these factors between HFD and CR. Overall, we found that only 1.2% (381

regions) of RXRα binding sites were differential between the diets, whereas a greater percentage (9.3%, 1,201 regions) of identified PPARα binding sites showed some evidence for differential binding. However, only a small portion of these differential sites mapped near genes found to be differentially expressed between HFD and CR, though many of these genes do indeed possess at least some binding evidence for these factors within or near their boundaries. We highlighted *Abcc4 and Cyp7a1* as examples of genes that change in expression between the diets and that also possess a differential binding region for PPARα nearby.

While PPARα is a well-established regulator of lipid metabolism in the liver [273], we noted extensive binding for this factor near genes involved in glucose metabolism. Prior studies of PPARα mutant mice [277, 284], induced activation of PPARα in mice [275, 285], and others [276, 286] have also suggested a role for PPARα-dependent regulation of carbohydrate metabolism. Here, we found evidence for PPARα binding near many genes specifically involved in the glycolysis/gluconeogenesis pathway (9 of 14 genes in the canonical pathway analyzed), many of which are sensitive to PPARα agonist treatment according to prior data [275] and five of which are altered in expression in response to HFD and/or CR according to our RNA-Seq data. To further test the role of PPARα in regulating glucose metabolism, we performed *in vitro* experiments in mouse primary hepatocytes and *in vivo* experiments in mice following fenofibrate treatment. We found that activation of PPARα by fenofibrate enhanced lactate production in the presence of glucose, but decreased glucose in the presence of lactate as a fuel. These results suggest a role for PPARα in enhancing anaerobic glycolysis in the liver. To further test these results, we showed that fenofibrate treatment reduces oxygen consumption rates in hepatocytes. We also found that fenofibrate treatment reduces the expression of the genes *Fbp1*, *Gck*, and *Pklr in vivo*, all of which are novel PPARα-regulated genes identified in this study that are involved in glucose metabolism and contain clear binding sites for PPARα near their transcription start sites. Indeed, evidence for a regulatory role of PPARα on *Gck* expression is somewhat contradictory from previous studies [277]. Fibrate has been shown to decrease its expression in mouse (as we see here), though the PPARα agonist WY14643 does not have the same effect. Moreover, rats possess a PPAR response element (PPRE) near this gene that is activated by LXRα/RXRα and PPARγ/RXRα in luciferase assays, though the role of PPARα in such studies has not been elucidated. Here, we show that PPARα indeed binds near the liver

promoter of *Gck* and that the expression of this gene is sensitive to *in vivo* fenofibrate treatment. Overall, our results strongly suggest a role for PPARα in regulating glucose metabolism, in particular anaerobic glycolysis.

4.4. METHODS

4.4.1. Animals and treatments

Calorie restricted male C57BL/6J mice (5 months of age, 40% calorie restriction [287], 13.7% calories from fat) were obtained from Charles River Laboratories. Additional male C57BL/6J mice (stock number 000664, Jackson Labs, Bar Harbor, ME) were fed a standard normal (chow) diet (Prolab Isopro RMH 3000, LabDiet, St. Louis, MO, 14.3% calories from fat) or a high fat diet (HFD) (TD.93075; Harlan Laboratories, South Easton, MA, 54.8% calories from fat) for a period of 16 weeks with free access to food and water. All mice used in this study were housed in a facility accredited by the American Association for Laboratory Animal Care (AALAC). Calorie restricted mice were acclimated in the same animal facility as the chow and HFD mice prior to euthanasia. All experiments were carried out in accordance with guidelines for the use of laboratory animals and were approved by the Institutional Animal Care and Use Committees (IACUC) of University of Massachusetts Medical School and Massachusetts Institute of Technology.

Glucose tolerance tests were performed by intraperitoneal injection of mice with glucose (1 g/kg). Insulin tolerance tests were performed by intraperitoneal injection of mice with insulin (0.5 U/kg). Pyruvate tolerance tests were performed by intraperitoneal injection of mice with pyruvate (1 g/kg). Assays were performed using methods described previously [224]**.**

We also injected 8 week old C57BL/6 male mice intraperitoneally with the fenofibrate (100 mg/kg), the PPARα antagonist GW6471 (10 mg/kg), or with vehicle (DMSO/Solutol HS15/Sterile water) (10:15:75) three times a week over a two week period.

4.4.2. RNA-Seq

Total RNA was extracted from the livers of mice (three per dietary condition) fasted overnight using the RNeasy Plus Mini kit (Qiagen, Valencia, CA). mRNA was isolated from DNA-free total RNA using an Illumina mRNA Purification Kit (Illumina, San-Diego, CA). The cDNA library was size-fractionated via gel electrophoresis by cutting a narrow slice (~2 mm, +/- 25bp) of the cDNA lane centered at the 300bp marker. cDNA from the gel slice was extracted using the Qiagen PCR mini elute kit (Qiagen). The sample was then amplified by PCR using the paired-end primers and amplification reagents supplied with the Illumina ChIP-Seq genomic DNA prep kit. The amplified product was purified using a Qiagen PCR mini elute kit (Qiagen). The library was then used to build clusters on the Illumina flow cell according to the manufacturer's protocol.

Following sequencing, the raw paired-end reads were aligned to known mouse RefSeq gene transcripts obtained from the UCSC table browser [122] (accessed on May 19, 2016) and the mouse genome (build mm9) with the splice junction-aware short-read alignment tool TopHat (version 2.1.0) [225]. We restricted TopHat to only align to known transcript splice junctions. We used the Bioconductor package conditional quantile normalization (CQN, version 1.6.0) [226] to remove systematic biases due to GC-content and gene length coverage and used DESeq2 (version 1.0.18) [227] 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).

4.4.3. Clustering and enrichment analyses

All hierarchical clustering was performed with the *clustergram* function in Matlab with Euclidean distance and average linkage. For enrichment analyses, we used custom Matlab code implementing the hypergeometric distribution for enrichment p-value calculations and used the Benjamini-Hochberg FDR procedure to correct for multiple hypotheses [257].

4.4.4. Microarray analysis

Raw CEL files from a published microarray study were obtained from the Gene Expression Omnibus, accession number GSE12147 [275]. This included data from male C57BL/6 mice treated with several selective PPARα agonists for 24hr or 5 days at 1mg/kg/day or water (vehicle) as control. Samples were background adjusted and normalized using the Bioconductor package *gcrma* and tested for differential expression between conditions using *limma* [288] in R.

4.4.5. DNase-Seq

We performed DNase-Seq on livers from mice fed CD, HFD, or CR according to a previously described protocol [289]. Briefly, liver nuclei were isolated from a pool of 3-4 mice using sucrose based buffer and digested with DNaseI (Promega, Madison, WI). The chromatin was incubated overnight with Proteinase K (Life technologies, Grand Island, NY) at 55^{o}C. DNA was extracted using phenol chloroform and small DNA fragments were isolated using a sucrose gradient ultracentrifugation followed by a gel size selection step. The DNA fragments were subjected to library preparation and sequencing according to the Illumina protocol.

Sites of DNase cleavage are identified as the 5' ends of the sequenced short reads from the DNase-Seq assay. We used the GPS algorithm [290] to identify regions of enriched cleavage compared to a control DNase-Seq assay performed on naked genomic DNA (proteins stripped from the chromatin by phenol-chloroform extraction). GPS builds a probabilistic mixture model to predict the most likely positions of binding events at single-base resolution, requiring an empirical spatial distribution of DNase reads around a typical binding event to build its event detection model. To build the empirical distribution, we identified binding regions from PPARα and RXRα ChIP-Seq data in the same condition, centered in on regions containing known motifs for the protein in question, and summed the DNase read distribution at every base pair in a 600 base pair window around these binding sites. We also performed pairwise comparisons between conditions by submitting both DNase datasets to GPS in multiple condition mode.

4.4.6. Motif analyses

For DNase hypersensitive sites, we took a 100 bp window around the single base GPS-identified sites for calculation of CpG content and motif matching. We calculated normalized CpG content of sequences using [291, 292]:

$$\text{NormalizedCpG} = \frac{\text{ObservedCpGs}}{(\text{ExpectedCpGs} \mid \text{GC content})} = \frac{\text{ObservedCpGs}}{(\text{GC content}/2)^2}$$

and divided sequences into low (<0.5) and high (>0.5) CpG content sets based on the bimodality of the empirical CpG content distribution obtained.

For motif analyses, we used a set of 1,588 DNA-binding motifs annotated to human and mouse transcriptional regulatory proteins from release 2013.3 of TRANSFAC® [293], represented as position-specific scoring matrices (PSSMs). All motifs used were of sufficient total information content (>8 total bits). We extracted the underlying genomic sequences from DNase hypersensitive regions and used TAMO [235] to store the motif PSSMs, read in 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 region, 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. Regions with no matches to a given motif were given a score of zero. We also computed motif match scores for sets of equally-sized, GC-content matched background sequences obtained by randomly sampling regions from the mm9 genome.

We used a hypergeometric test to determine enrichment of a motif in the sets of foreground sequences (i.e. DNase regions) compared to matching random background sequences. For such tests, we counted, for a given motif, the number of motif matches in both the foreground and background sets of sequences and compared these values to one another. As many of the motif models are redundant, we used affinity propagation [244] to cluster the motifs, using the pairwise Kullback-Leibler divergence as the similarity metric and a self-similarity parameter of -0.4. This procedure created 284 motif clusters. We post-clustered the motif enrichment results, retaining

the result from the most significantly enriched motif in each cluster, and corrected the raw p-values with the Benjamini-Hochberg procedure.

4.4.7. ChIP-Seq

Following overnight fasting, mice were anaesthetized and the liver was processed as previously described [294]. ChIP experiments were performed on two livers per condition (biological replicates) using antibodies against RXRα (sc-153x, Santa Cruz Biotechnology, Santa Cruz, CA) or PPARα (MAB3890, Millipore, Billerica, MA). We fragmented chromatin with a Covaris S2 sonication machine (Covaris, Woburn, MA) to obtain fragments ranging from 500 to 1000 base pairs. 5 µg of antibody or IgG was incubated with beads for 6 hours before incubating with sonicated chromatin overnight. We then washed the beads, eluted the chromatin, reversed crosslinks for 2 hours, and treated samples with RNase and Proteinase K. We purified the DNA and constructed sequencing libraries using the DNA Sample Kit (Part# 0801-0303, Illumina, San Diego, CA) according to the manufacturer's instructions. The samples were sequenced on an Illumina GAII/HiSeq sequencing platform and the resulting short reads were aligned against the mm9 reference mouse genome using Bowtie (version 0.12.7) [295]. Enriched genomic regions were identified by MACS (version 1.4) [296] using an IgG control and the resulting peaks were filtered to have an enrichment p-value of <1e-10. Overlapping peaks between RXRα and PPARα ChIP-Seq datasets were restricted to those whose summits mapped within ±100 bp. Transcription factor binding motifs from the TRANSFAC database were used with the THEME software package [125] to find enriched motifs in the DNA sequences under the filtered ChIP peaks. For ChIP-Seq read pileup visualizations, we concatenated the aligned sequence reads from biological replicates for each factor in each condition, extracted reads mapping within the specified windows (allowing for only two reads mapping to the exact same location to minimize PCR biases), extended each read in the 3' direction to a full length of 200 bp, summed the number of extended reads overlapping each base pair within the window, normalized the read count levels to account for sequencing depth differences between samples, and smoothed the read profiles using a moving average filter (120 bp rate). Thus, read pileup axes in figures 4-3 and 4-5 refer to these concatenated, extended, normalized, and smoothed read profiles.

4.4.8. Primary hepatocytes

We isolated mouse primary hepatocytes with a modified 2-step perfusion method [297] that uses Liver Perfusion Media and Liver Digest Buffer (Invitrogen) [298]. We seeded cells on plates (pre-coated [1 h] rat tail collagen I [BD Biosciences]) in DMEM supplemented with 4.5 g/L glucose, 10% FBS, 0.2% BSA, 2 mM sodium pyruvate, 2 mM glutamine, 1 μM dexamethasone, 100 nM insulin and 1% penicillin /streptomycin. After attachment (2 h), the medium was removed and the hepatocytes were incubated (22 h) in maintenance medium (DMEM supplemented with 4.5g/L glucose, 0.2% BSA, 2 mM sodium pyruvate, 2 mM glutamine, 0.1 μM dexamethasone, 1 nM insulin and 1% penicillin/ streptomycin). In some cases, we incubated hepatocytes (16 h) with fenofibrate (100 μM, Sigma). The drugs were dissolved in DMSO; control studies were performed by addition of vehicle (DMSO) alone.

We evaluated glucose production by incubating $5.5x10^5$ primary hepatocytes in collagen-coated 35 mm wells (6 well plates) with M199 media (Invitrogen) supplemented with 0.5% BSA and 1% penicillin/streptomycin for 18 hours. Cells were then incubated in glucose/glutamine/phenol red-free DMEM (Sigma) supplemented with 3.7 g/L sodium bicarbonate, 2 mM lactate and 20 mM sodium pyruvate for the indicated times. Glucose production in the medium was assessed using the glucose (HK) assay kit (Sigma) and values were normalized to total hepatocyte protein.

We evaluated lactate production by incubating $5.5x10^5$ primary hepatocytes in collagen-coated 35 mm wells (6 well plates) with M199 media (Invitrogen) supplemented with 0.5% BSA and 1% penicillin/streptomycin for 18 hours. Cells were then incubated in glucose/glutamine/phenol red-free DMEM (Sigma) supplemented with 1.85 g NaCl, 0.2% BSA, 0.1 μM dexamethasone, 1 nM insulin and 138 mM glucose for the indicated times. Lactate production was measured in the medium using the reconstituted Lactate Reagent (Beckman Coulter) and values were normalized to total hepatocyte protein.

4.4.9. Oxygen consumption rates

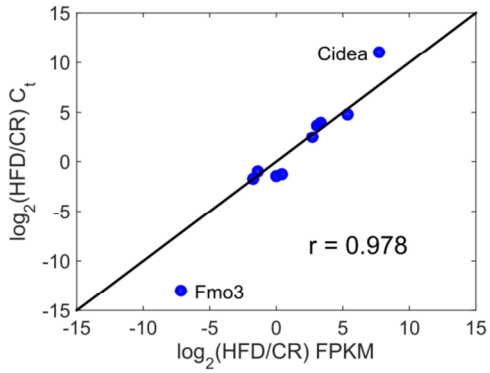
We quantified oxygen consumption rates (OCR) in primary hepatocytes using an XF24 Extracellular Flux Analyzer (Seahorse Bioscience, Billerica, MA) and XF assay kits to measure extracellular flux changes of oxygen and protons. Briefly, primary hepatocytes were plated ($4x10^4$ cells/well) in collagen-coated XF24-microplates (Seahorse Bioscience). After attachment (2 h), the hepatocytes were transferred to running medium (sodium bicarbonate-free DMEM supplemented with 4.5 g/L glucose, 0.2% BSA, 2% penicillin/streptomycin, 1 nM insulin and 0.1 µM dexamethasone) and incubated at 37°C in a humidified atmosphere without CO_2 supplementation. Baseline measurements were performed prior to the addition of substrates (1 g/L glucose, 200 µM palmitate-BSA, or 10 mM lactate/1 mM pyruvate) or inhibitors (1 µM oligomycin, 0.1 µM FCCP, or 100 nM rotenone). Mitochondrial oxygen consumption rates were calculated as the difference between the maximal respiratory rate (in the presence of FCCP) and the respiratory rate after addition of rotenone. Data obtained from 11 independent wells were examined for each condition.

4.4.10. Quantitative RT-PCR

The expression of mRNA was examined by quantitative PCR using a Quantstudio PCR machine (Life Technologies). TaqMan® assays were used to quantify *Acox1* (Mm01246834_m1), *Ehhadh* (Mm00619685_m1), *Fbp1* (Mm00490181_m1), *Gck* (Mm00439129_m1), *Pdk4* (Mm01166879_m1), and *Pklr* (Mm00443090_m1). The relative mRNA expression was normalized by measurement of the amount of *18S* RNA in each sample using TaqMan© assays (catalog number 4308329; Life Technologies).

4.5. SUPPLEMENTARY FIGURES

A



B

Sample	Total paired-end reads	% Aligned reads	% Proper pair alignments	% Aligned uniquely
CD-1	37643030	97.5	94.1	79.2
CD-2	33414197	96.6	92.4	79.6
CD-3	35945949	96.6	92	79.2
HFD-1	36262557	97.2	93.8	81.3
HFD-2	37447506	96.5	92.8	77.3
HFD-3	25037298	96.6	93.5	84.3
CR-1	39189223	96.6	92.5	85.5
CR-2	19858320	96.9	93.7	85
CR-3	30429450	97.7	94.8	84.7

Figure 4-S1. qPCR validation of CR versus HFD gene expression changes and RNA-Seq sequence read alignment statistics. (A) qPCR validation of gene expression level changes between CR and HFD for genes *Alb, Apoa1, Apoa4, Cidea, Egr1, Fmo3, Fos, Il1rn, Rps14,* and *Sirt3*. (**B**) Total obtained reads, alignment percentages, proper paired mapping percentages, and the percentage of uniquely aligned reads for paired-end reads from CD, HFD, and CR RNA-Seq samples.

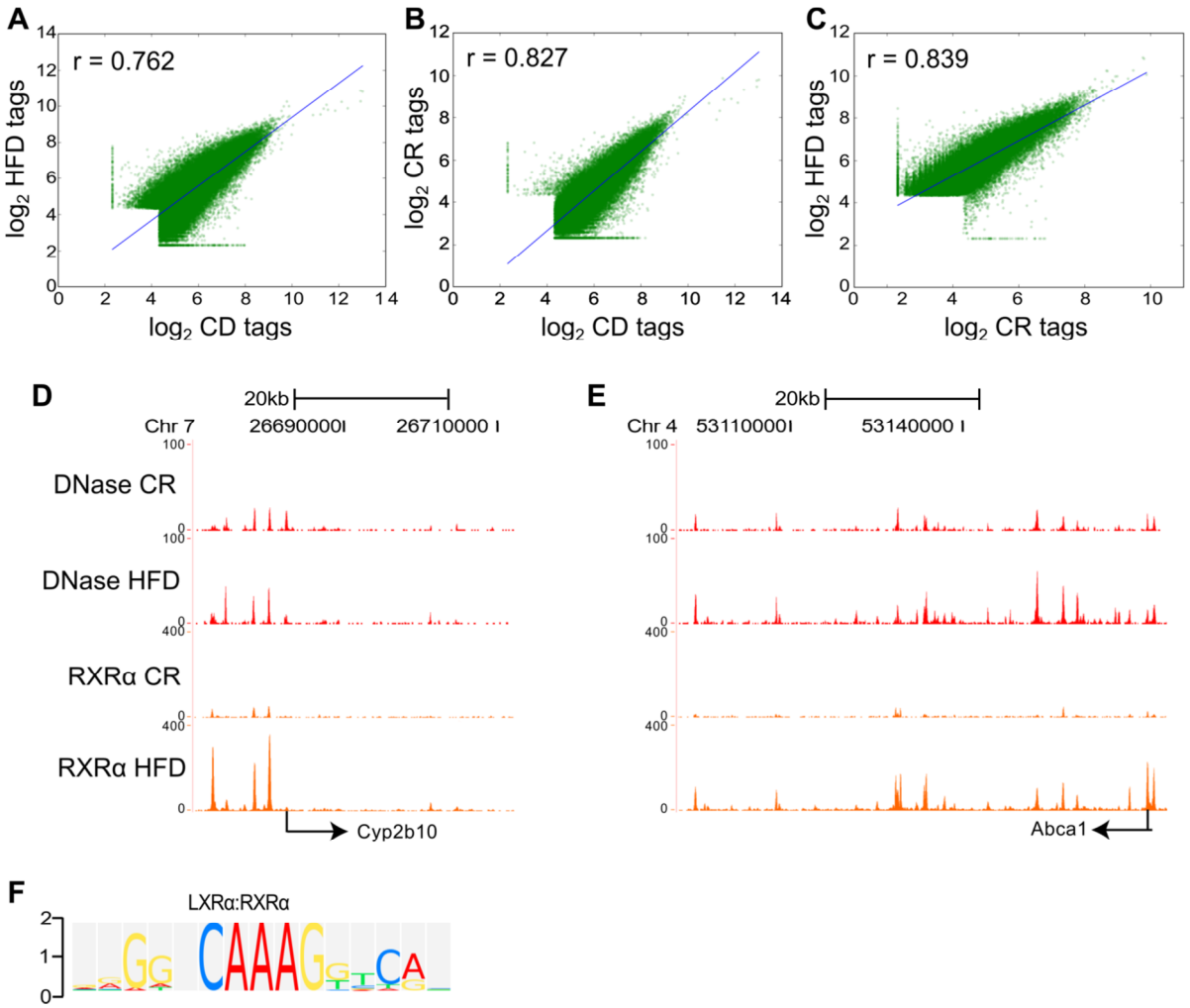


Figure 4-S2. DNase-Seq dataset correlations and example binding profiles. (A-C) Correlation plots between read counts (tags) from CD, HFD, and CR DNase-Seq datasets: HFD versus CD (A), CR versus CD (B), and HFD versus CR (C). Correlation values are for Pearson correlation coefficients. (D-E) Example read pileup tracks from CR and HFD DNase-Seq datasets near genes *Cyp2b10* (D) and *Abca1* (E) which are known to contain LXRα:RXRα binding sites. Bottom tracks show RXRα profiles from CR and HFD ChIP-Seq samples, confirming binding sites in these hypersensitive regions for expected factors. (F) Motif logo for LXRα:RXRα DNA-binding preference that is enriched in DNase-Seq regions.

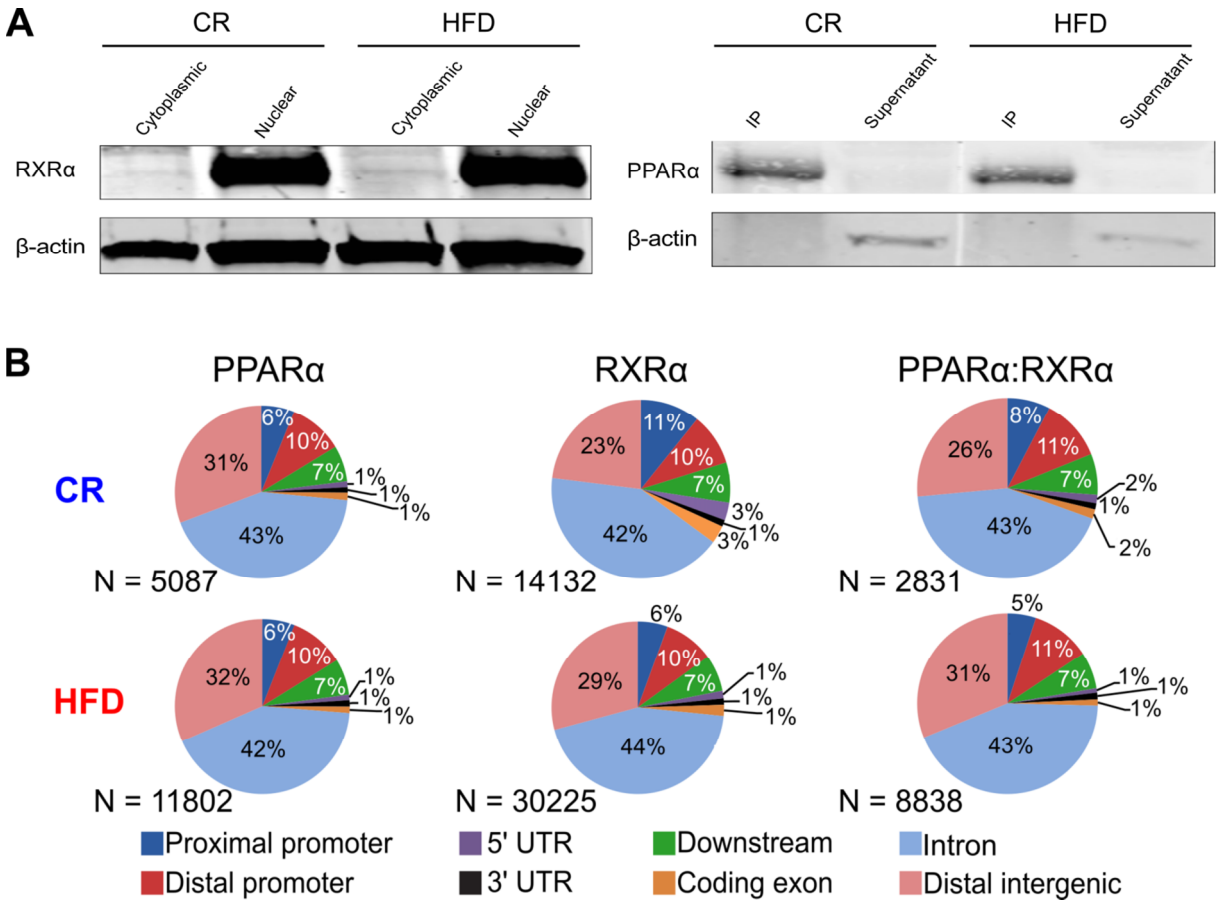


Figure 4-S3. Validation of PPARα and RXRα antibodies and binding locations from ChIP-Seq studies. (A) (Left) Nuclear or cytoplasmic fractions of homogenized CR and HFD livers were assayed by Western blot using anti-RXRα or β-actin primary antibodies. (Right) Whole-cell lysates from CR and HFD livers were immunoprecipitated with PPARα antibody; IP and supernatant were immunoblotted with anti PPARα or β-actin antibodies. (B) Genome-wide binding locations of PPARα, RXRα, and overlapping PPARα:RXRα ChIP-Seq peaks with total enrichment region numbers. Regions were mapped near genes according to: proximal promoters – within 200 bp of gene TSS; distal promoters – within 5 kb upstream of gene; downstream – within 5 kb downstream of gene end; introns, exons, 5' UTR, and 3' UTR – if region intersected one of these features; and distal intergenic – outside 5 kb window around gene.

CHAPTER 5

INFERRING MICRO RNA-MEDIATED REGULATORY ACTIONS FOLLOWING OBESITY-INDUCED HEPATIC INSULIN RESISTANCE

Micro RNAs (miRNAs) are a small (~22 nucleotides) class of RNA species that target and regulate mRNAs post-transcriptionally, affecting a wide variety of biological processes that are relevant to disease. These molecules have been examined in the context of cancers, neurodegenerative diseases, and metabolic diseases. Small RNA-Seq is a method to comprehensively profile the full expression landscape of miRNAs in cells and tissues. In this work, we performed small RNA-Seq on liver samples collected from mice fed a standard laboratory chow diet (CD), a 6 week high-fat diet (HFD), and a 16 week HFD to analyze miRNA expression profiles in the context of hepatic insulin resistance. We found that HFD progressively alters the expression of miRNAs in the liver. We integrated these data with mRNA-Seq and histone modification ChIP-Seq data collected from mice fed these same diets to identify potential regulatory roles conferred by these miRNAs. Specifically, we used an enrichment scheme that considered overrepresentation of mRNAs in the pool of differentially regulated genes by HFD that are targets of each miRNA and used a network modeling algorithm that incorporated miRNA, mRNA, and epigenetic data to specifically probe miRNA-transcription factor interactions. Both methods prioritized miRNAs with both known and potentially novel regulatory roles in the context of hepatic insulin resistance that can be readily examined with additional targeted experiments.

Anthony R. Soltis, Xiaofeng Xin, Heather M. Sweeney, Yoonjeong Cha, Yoon Sing Yap, Bryan J. Matthews, Norman J. Kennedy, Roger J. Davis, Ernest Fraenkel

Author contributions: A.R.S. wrote code to analyze small RNA-Seq data, performed mRNA-Seq analysis, ran code to infer miRNA-gene targets, determined transcription factor-gene interactions, performed miRNA-gene target enrichment analyses, and compiled results and wrote text; X.X. performed small RNA-Seq, mRNA-Seq, and ChIP-Seq experiments, with assistance from Y.S.Y. and B.J.M.; A.R.S. and X.X. analyzed small RNA-Seq results; H.M.S. performed SAMNet modeling; Y.C. contributed to SAMNet modeling; N.J.K. performed mouse experiments and prepared liver samples; A.R.S., X.X., N.J.K., R.J.D., and E.F. conceived of and designed the work; R.J.D. and E.F. supervised the project.

Acknowledgments: We acknowledge members of the MIT BioMicro Center for assistance with sequencing data collection. This work was supported by R24 DK-090963 (E.F. and R.J.D.), R01 NS-089076 (E.F.), R01 DK107220 (R.J.D.), and used computing resources funded by the National Science Foundation under Award No. DB1-0821391 and sequencing support from NIH (P30-ES002109).

5.1. INTRODUCTION

Micro RNAs (miRNAs) are a class of small (~22 nucleotide) RNA species that target and regulate mRNAs post-transcriptionally [299, 300]. In mammals, greater than 60% of protein-coding mRNAs possess at least one conserved miRNA target site in their 3' UTRs [301], while the existence of non-conserved and non-3' UTR target sites (e.g. in coding exons) points to the potential for miRNAs to regulate a substantial majority of protein-coding genes [300]. miRNAs exert their regulatory activities by pairing with target sequences on mRNAs and promoting either translational inhibition or mRNA degradation, the latter of which appears to be the major mechanism [302, 303]. miRNA biogenesis and activation occur through a multi-step process: primary miRNAs (pri-miRNAs) are transcribed from the introns of coding and non-coding transcripts (though some derive from exons); the nuclear RNase III Drosha in complex with DGCR8 crops a ~65 base pair stem-loop hairpin (called a pre-miRNA) from the pri-miRNA;

exportins assist in translocation of the pre-miRNA from the nucleus to the cytosol; the endonuclease Dicer cleaves pre-miRNAs to produce a small RNA duplex; the miRNA duplex is loaded onto an AGO protein to form an RNA-induced silencing complex (RISC); and the passenger strand of the duplex is removed to form a mature RISC that can initiate the miRNA's regulatory activities [300].

Critical regulatory roles for miRNAs have been identified in the contexts of numerous biological processes and diseases, including development [304], differentiation [305], apoptosis [306], cancers [307], cardiovascular diseases [308], neurodegenerative diseases [309], and autoimmune diseases [310]. Relevant to this work, miRNAs in liver and other peripheral tissues have also been associated with a variety of processes related to obesity and insulin resistance [311, 312]. These include miR-122 [313], miR-802 [314], miR-103 and miR-107 [315], miR-33 [316], miR-143 [317], and miR-181 [318], all of which likely regulate a host of targets across many biological processes. While many regulatory roles for hepatic miRNAs in the context of insulin resistance have been identified to-date, the sheer number of potential regulated targets warrants further exploration into mechanisms that may prove exploitable as therapeutic interventions.

Small RNA-Seq is a powerful experimental tool that can comprehensively quantify full miRNA expression landscapes in cells and tissues. While it is fairly straightforward to arrive at lists of differentially expressed small RNAs in a system of interest, interpreting and prioritizing such results remains challenging. Collection of complementary data, including transcriptome-wide mRNA expression profiling, can drastically improve analyses of miRNA-mediated activities relevant to disease.

The identification of target mRNAs, along with analysis of miRNA expression profiles themselves, is a crucial step to understanding miRNA-mediated actions. miRNAs recognize complementary sequences on target mRNAs with nucleotides 2-7, called the miRNA seed, and downstream sites can additionally aid target recognition [319]. Both experimental and computational approaches can be used to identify miRNA targets. High-throughput experimental methods, including AGO CLIP-Seq [320] and CLASH [321], can directly identify tissue and condition-specific targets through isolation, sequencing, and analysis of AGO-associated RNA

duplexes. Computational approaches to target identification use a variety of features for prediction, including seed pairing, conservation, site number, and 3' supplementary pairing. Such algorithms include TargetScan [301], DIANA-microT [322], and PITA [323]. These datasets and tools are immensely useful for generating hypotheses regarding relevant miRNA-mRNA target interactions.

In this work, we performed small RNA-Seq on the livers of mice fed chow diets (CD), short-term (6 week) high-fat diets (HFD), or long-term (16 week) HFD. We additionally incorporated mRNA-Seq and histone modification ChIP-Seq data from these animals with computational methods. Specifically, we developed a scheme that ranked differentially expressed miRNAs by the enrichment of predicted target mRNAs (determined by TargetScan) present in the pool of genes differentially regulated following HFD. We also utilized a network modeling algorithm (simultaneous analysis of multiple networks or SAMNet [162, 324]) to directly test hypotheses regarding miRNA regulatory influences on transcription factor expression levels and subsequent downstream gene expression. Both of these methods identified miRNAs that are known to regulate processes related to hepatic insulin resistance, while also identifying new species that may be relevant to this metabolic condition.

5.2. RESULTS

5.2.1. HFD induces progressive dysregulation of hepatic miRNA expression profiles

We fed male C57BL/6J mice a CD, a 6 week HFD, or a 16 week HFD and profiled their hepatic transcriptomes (via mRNA-Seq) and epigenomes (via ChIP-Seq experiments for the histone modifications H3K4me3, H3K27Ac, and H3K27me1). Major findings from analyses of these datasets are reported in ***Chapter 2*** of this document. Briefly, we found that HFD feeding progressively induces widespread changes in hepatic gene expression, while producing minimal detectible differences in histone modification profiles and amounts. We did utilize, however, these epigenomic datasets, along with DNA binding motif data, to identify regulatory factors that likely influence the observed transcriptional changes.

We also specifically profiled hepatic miRNA expression profiles in these mice using small RNA-Seq. We built a custom analysis pipeline to process the obtained raw sequencing reads, to align reads to known mouse miRNA hairpin sequences, and to quantify the expression levels of known mature miRNA species. We performed differential miRNA expression analyses on these datasets and found that 15 and 50 miRNAs are significantly altered by 6 and 16 week HFD compared to CD, respectively (**Figure 5-1**). Six miRNAs were altered in the same direction by both diets: up-regulated miR-674-3p, miR-34a-5p, miR-149-5p, and miR-532-5p and down-regulated miR-21-5p and miR-5117-5p. Among these, miR-21 activation in pancreatic β-cells reduces cell death through anti-apoptotic regulatory actions [325], but up-regulation in response to high glucose in kidney is associated with diabetic renal injury and pathology [326, 327]. miRNAs specifically altered by 6 week HFD include miR-125a-5p and miR-30c-5p, both of which are down-regulated.

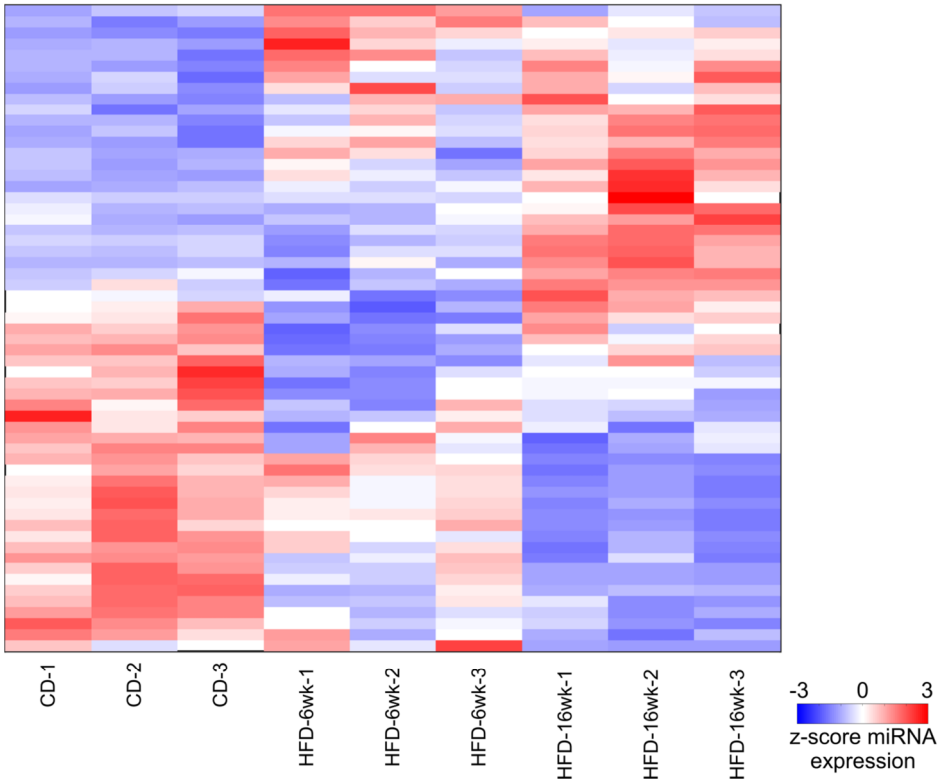


Figure 5-1. Differential miRNAs in 6 and 16 week HFD livers. We found 15 and 50 miRNAs differentially expressed by 6 and 16 week HFD compared to CD. The heatmap shows hierarchically clustered z-scored (miRNA-wise) expression values in replicate livers across the three conditions.

Feeding a HFD for 16 weeks expanded the pool of dysregulated miRNAs. Among these additional miRNAs are up-regulated miR-802-5p, miR-185-5p, and miR107-3p and down-regulated miR-379-5p and miR-182-5p. miR-802 is known to be up-regulated in the livers of obese mice and humans and is thought to impair glucose tolerance and insulin signaling via regulation of *Hnf1b* [314]. miR-185 has been shown to regulate lipid and cholesterol metabolic processes in hepatic cells [328, 329] and miR-107/103 up-regulation in obese mouse livers can impair glucose homeostasis [315]. Also, hepatocyte-specific reduction of miR-379 reduces very low-density lipoprotein levels in mice [330]. Thus, HFD induces alterations in miRNA expression profiles that, consistent with our observations for mRNAs, progressively diverge from CD expression levels as HFD feeding duration increases. We also identified miRNAs differentially expressed in our data that are consistent with prior, focused studies on the roles of these species in related metabolic processes.

5.2.2. miRNA-mRNA integration and target enrichment analysis prioritize miRNAs

We next sought to prioritize miRNAs for potential follow-up studies, focusing on their roles in alleviating or exacerbating the effects of obesity on hepatic insulin resistance. To do this, we devised an approach whereby we ranked miRNAs by statistical enrichment for target mRNAs differentially expressed by HFD in our data (**Figure 5-2**). We used TargetScan to generate miRNA-mRNA target predictions and queried our set of differential mRNAs for target matches to each miRNA. As miRNAs mainly serve inhibitory functions through enhancement of mRNA degradation and translational repression, we focused on miRNA-gene target pairs where the expression fold-changes of the two species induced by HFD were anti-correlated (i.e. if HFD increased a miRNA's expression, a potential target must have decreased). We then used the hypergeometric distribution to compute enrichment statistics for overrepresentation of predicted targets in the pool of differentially expressed genes for each miRNA.

We performed this enrichment analysis on the differentially expressed miRNAs between CD and both HFDs (**Table 5-1**). We found that miR-674-3p, miR-34a-5p, miR-149-5p, and miR-532-5p, all of which are up-regulated by both HFD conditions, are strongly enriched for differential

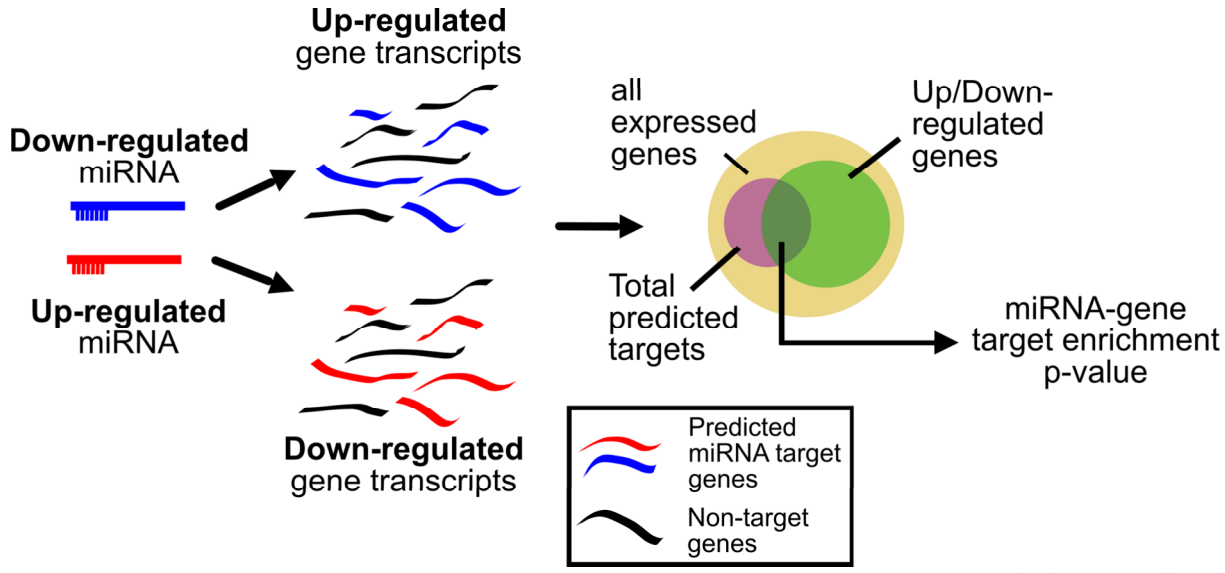


Figure 5-2. Schematic of miRNA-mRNA target enrichment analysis. Up and down-regulated miRNAs were matched to differentially expressed mRNAs that are predicted targets and anti-correlated in terms of expression change following HFD. We used the hypergeometric test to search for miRNAs with significant target enrichments in differential genes.

target genes in both analyses. In particular, miR-34a serum levels are up-regulated in patients with non-alcoholic fatty liver disease [331] and it has been shown to regulate retinoid X receptor α (RXRα), Krüppel-like factor 4 (KLF4), and Sirtuin 1 (SIRT1) mRNA in liver [306, 332, 333]. In 16 week HFD, up-regulated miR-152-3p appears at the top of the enrichment list. This miRNA is significantly elevated in obese patients [334] and can modulate the Wnt signaling pathway through targeting of DNA methyltransferase 1 [335]. Mice overexpressing miR-378, which is also up-regulated by 16 week HFD and highly ranked, display hepatic insulin resistance via targeting of the catalytic subunit (p110α) of phosphoinositide 3-kinase (PI3K) [336].

While our analysis indeed identified miRNAs with known roles in regulating processes relevant to hepatic insulin resistance, we also identified several high-ranking species whose roles are not well established. In particular, miR-1839-5p is substantially expressed and up-regulated by HFD in our data and is highly ranked by our enrichment scheme. This miRNA is a non-canonical small RNA, meaning it bypasses the typical Drosha processing pathway but is still cleaved by Dicer for its biogenesis [337], whose predicted target genes include *Crat* and *Pdk4*. miRNAs miR-149-5p, miR-455-5p, and miR-532-5p are also among the group of high-ranking targets where little is known with respect to their potential roles in regulating hepatic insulin resistance.

CD vs. 6wk HFD miRNAs		
miRNA	**% DEG targets**	**q-value**
mmu-miR-149-5p	43.2%	2.83E-54
mmu-miR-674-3p	36.9%	3.84E-49
mmu-miR-34a-5p	37.8%	2.29E-39
mmu-miR-27a-3p	31.5%	3.22E-39
mmu-miR-378-5p	25.0%	5.33E-28
mmu-miR-125a-5p	29.7%	2.48E-25
mmu-miR-145-5p	23.7%	4.04E-19
mmu-miR-19b-3p	19.8%	2.01E-18
mmu-miR-365-3p	16.5%	1.36E-15
mmu-miR-7a-5p	19.7%	2.40E-14
mmu-miR-532-5p	16.6%	1.22E-13
mmu-miR-21-5p	10.3%	2.32E-10
mmu-miR-5117-5p	8.8%	3.29E-07
mmu-miR-30c-5p	11.5%	4.59E-06
mmu-miR-122-3p	4.0%	0.0185798

CD vs. 16wk HFD miRNAs		
miRNA	**% DEG targets**	**q-value**
mmu-miR-152-3p	32.9%	4.62E-65
mmu-miR-34a-5p	36.1%	7.78E-62
mmu-miR-674-3p	31.8%	1.30E-58
mmu-miR-107-3p	31.7%	1.70E-56
mmu-miR-149-5p	34.8%	1.85E-51
mmu-miR-222-3p	23.0%	9.71E-46
mmu-miR-351-5p	31.0%	9.19E-45
mmu-miR-1839-5p	20.5%	7.96E-40
mmu-miR-378-3p	25.1%	5.05E-39
mmu-miR-185-5p	30.2%	5.16E-38
mmu-miR-98-5p	21.7%	6.87E-38
mmu-miR-148b-5p	15.6%	2.15E-36
mmu-miR-455-5p	20.2%	1.03E-35
mmu-miR-582-3p	17.5%	4.92E-33
mmu-miR-1843b-5p	23.9%	4.28E-31
mmu-miR-532-5p	18.1%	4.16E-29
mmu-miR-676-5p	15.7%	6.52E-29
mmu-miR-152-5p	17.3%	8.84E-29
mmu-miR-501-3p	16.9%	1.24E-28
mmu-miR-802-5p	15.5%	1.55E-28

Table 5-1. miRNA-mRNA target enrichment results. Results from enrichment analyses performed on differential miRNAs in 6 week (left) and 16 week (right) HFD. The second column reports the percentage of anti-correlated differential mRNAs that are predicted targets of the miRNAs and the third column lists the FDR-corrected enrichment p-values. Note that we only show the top 20 most significant enrichments for the CD vs. 16 week HFD analysis.

Thus, our enrichment analysis provides a basis from which candidate miRNAs can be identified for follow-up study based on potential gene targeting.

5.2.3. Integrative modeling of miRNA, mRNA, and epigenomic data reveals miRNA-regulated transcriptional networks

Our integrated miRNA-mRNA target enrichment analysis was indeed able to prioritize miRNAs based on overrepresentation of differentially expressed predicted targets. However, each miRNA still potentially regulates hundreds of mRNA species; therefore it is difficult to assess which interactions are most relevant to disease. We can reduce this search space by focusing on particular types of miRNA targeting interactions. In this vein, miRNAs may actually exert profound effects through their regulation of mRNAs encoding transcriptional regulators. Such

regulation can amplify the effects of altered miRNA activities through secondary alteration of downstream transcription mediated by these factors.

To search for such regulatory interactions in this context, we used the SAMNet network modeling tool [162] to generate a reduced graphical model of these potential regulatory interactions (**Figure 5-3**). SAMNet incorporates a flow-based algorithm that searches for the optimal path through a hierarchy of nodes and edges, starting at an artificial source node and collecting at a terminal sink node. This approach considers multiple types of evidence during the optimization and produces compact networks for simpler interpretation. In our formulation (based on [324]), nodes represent miRNAs, transcription factors, active epigenetic regions, and mRNAs, while edges were assigned weights according to miRNA expression changes, predicted miRNA-mRNA target scores (considering only anti-correlated relationships), mRNA expression changes in genes encoding transcription factors, predicted affinity scores for transcriptional

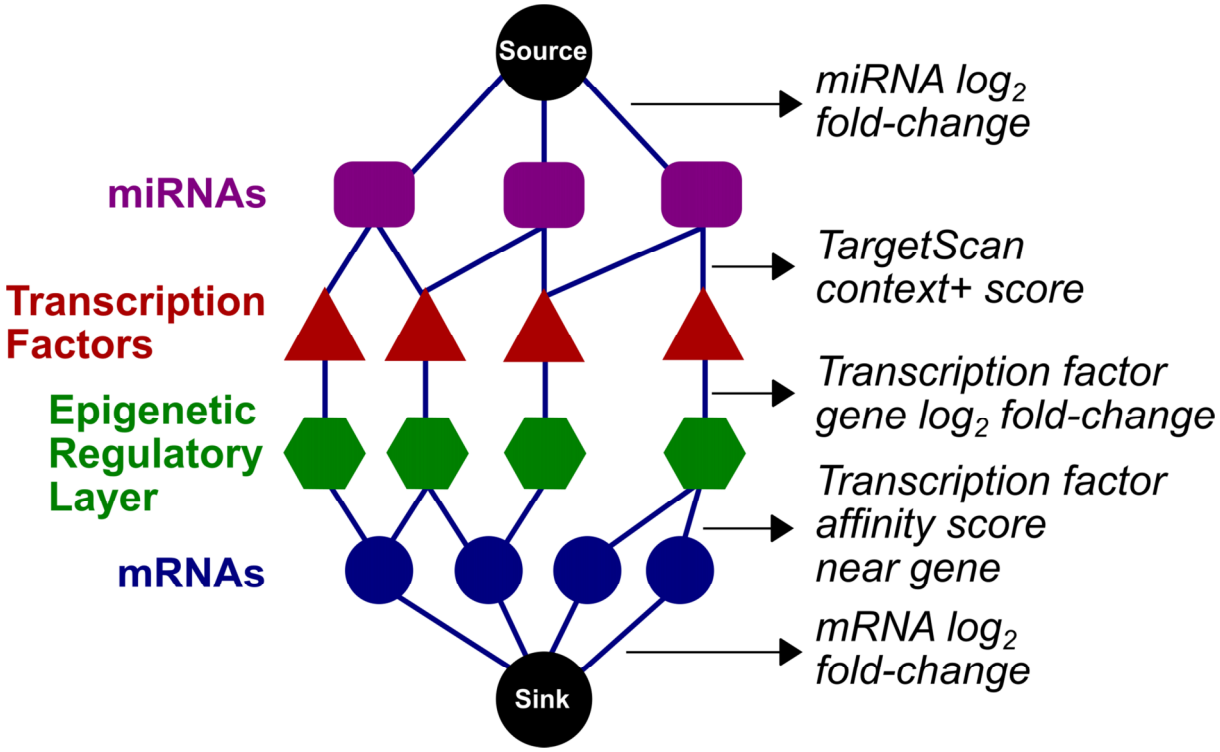


Figure 5-3. SAMNet formulation for miRNA regulatory network. SAMNet implements a flow-based algorithm that sends flow from source to sink through an intermediate regulatory layer. We included four regulatory layers (left) and used five types of evidence for edge weight assignments (right).

regulators near differentially expressed genes, and mRNA expression changes. We ran several iterations of the algorithm using different values of the γ tuning parameter, which effectively controls the size of the solution. The final network presented here was chosen because it is reasonably sized and solutions run at larger γ values were very similar to this.

The resulting network model run on our insulin resistance data is shown in **Figure 5-4**. We show only the algorithm-identified miRNA-transcription factor interactions for simplicity. The algorithm predicted that miR-34a and miR-149 may regulate several transcriptional regulators, including *E2f3*, *Sox12*, *Atf5*, *Hnf1b*, and *Nr6a1*, to affect transcriptional regulation in response to HFD. Among these predictions, *E2f3* is a validated target of miR-34a [338]. Both of these

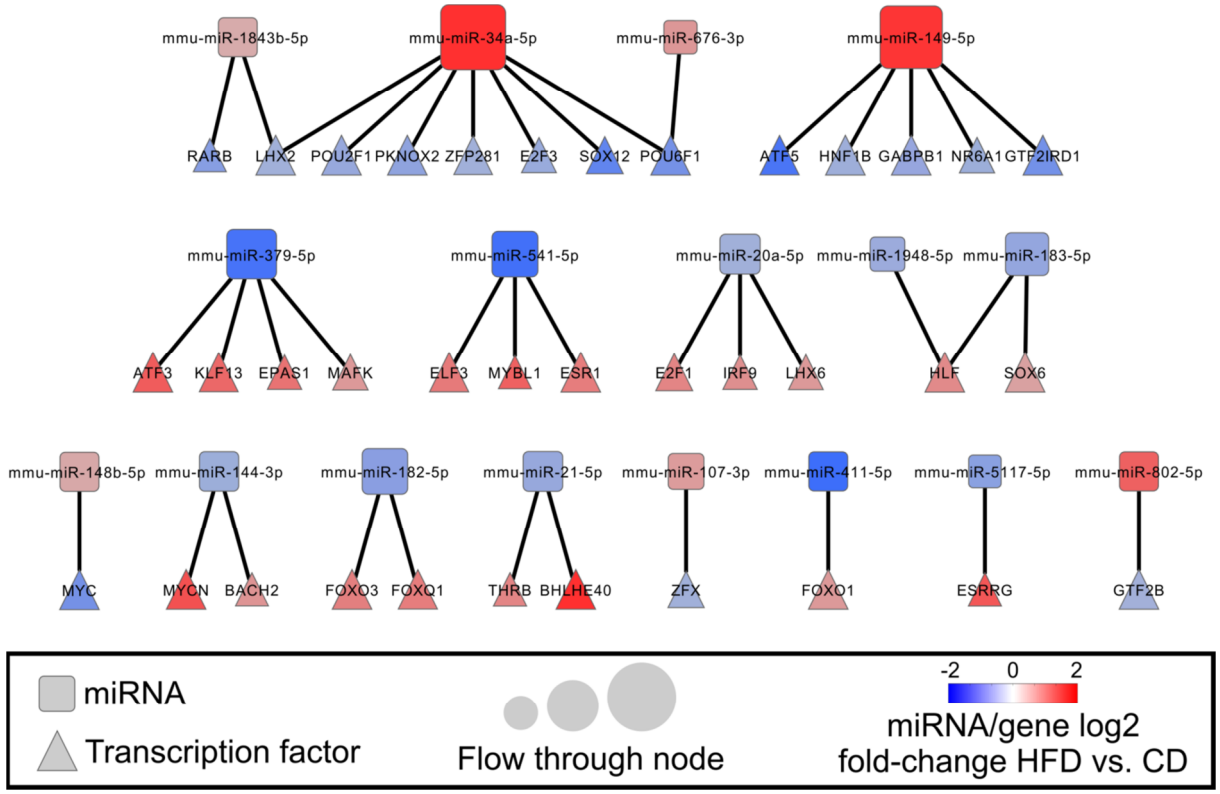


Figure 5-4. SAMNet model results. These results used a SAMNet tuning parameter (γ) equal to 16. Rounded squares and triangles represent miRNAs and transcription factors, respectively, node sizes reflect the relative amount of flow passing through each node of the graph, and the color scale reflects the level of differential expression between 16 week HFD and CD livers for miRNAs and transcription factors. We only show the miRNA-transcription factor regulatory layers here for simplicity.

miRNAs were also highly ranked by our target enrichment analysis. While *Hnf1b* has been shown to be regulated by miR-802 in obese livers [314], our model predicts that miR-149 may also play a role in regulating this factor. miR-802 itself is predicted to potentially play a role in regulating *Gtf2b*, which is a ubiquitous transcription factor required for initiating RNA polymerase II transcription. In alternate simulations (data not shown), miR-802 was also identified as a regulator of *Hnf1b*. The model also predicts several regulatory interactions for miR-379, one of which is targeting of *Atf3*, a transcription factor that is activated by stress signals and can induce liver dysfunction [339]. The network model additionally implicates miR-21, which is down-regulated by HFD, in regulating *Thrb* and *Bhlhe40*. A direct interaction between miR-21 and *Thrb* has been confirmed in other cellular contexts [340]; we show here evidence for this potential regulatory interaction in the liver. As a final example, miR-182 is predicted to regulate *Foxo3* and *Foxq1* mRNA. A regulatory interaction between miR-182 and *Foxo1* has been demonstrated in breast cancer cells [341] and helper T lymphocytes [342]. Our model implicates miR-411 in regulating *Foxo1* mRNA, an interaction that has recently been validated in lung cancer cells [343], while also highlighting roles for miR-182 in regulating additional forkhead family members. Thus, our network model provides known and novel predictions surrounding the roles of miRNAs in regulating liver transcription factors during insulin resistance.

5.3. DISCUSSION

In this work, we collected small RNA-Seq datasets to profile miRNA expression in CD, 6 week, and 16 week HFD mouse livers. Direct analysis of these datasets revealed that HFD progressively alters the expression landscape of hepatic miRNAs, with the short and long-term HFDs altering the levels of 15 and 50 miRNA species, respectively. Six miRNAs were significantly altered in the same direction by both diets, including miR-34a-5p, miR-149-5p, and miR-21-5p. We then used target predictions from TargetScan and mRNA-Seq data collected from these same conditions to prioritize interesting miRNAs for follow-up study. To do this, we used an enrichment scheme that queried the amounts of differential mRNAs that are predicted targets of each altered miRNA. This analysis pointed to miR-674-3p, miR-34a-5p, miR-149-5p, and miR-532-5p as relevant in both 6 and 16 week HFD. Several of these predictions correspond

to miRNAs with established roles in regulating biological processes relevant to insulin resistance; however, we did identify several high-ranking miRNAs whose roles, to our knowledge, are not established in these contexts, including miR-1839-5p, miR-149-5p, miR-455-5p, and miR-532-5p.

We also used a graphical modeling methodology to analyze miRNA regulation of transcription factor networks more explicitly. miRNA's may exert significant influence by participating in regulatory motifs with transcription factors and the downstream genes they regulate [344, 345]. We incorporated our miRNA expression, mRNA expression, and histone modification data using the SAMNet algorithm [162, 324] to model such interactions directly. The resulting network consisted of a wide array of potentially relevant miRNA-transcription factor interactions to hepatic insulin resistance. We noted that some of these interactions have been reported previously, including miR-34a with *E2f3* and miR-21 with *Thrb* [338, 340], though it appears that the majority of the model's regulatory paths have not been previously identified or considered in these contexts. Thus, our modeling efforts provided a substantial, yet tractable, set of miRNA-transcription factor interactions that may serve as additional regulatory mechanisms that either alleviate or promote the effects of obesity on hepatic insulin resistance.

Reliable determination of miRNA targets is crucial to the study of their regulatory functions. Here we used the TargetScan algorithm to computationally predict which mRNAs contain sites for miRNA regulation in their 3' UTRs [301, 319]. While TargetScan is generally a reliable tool for target site identification, high-throughput experimental approaches, including AGO CLIP-Seq [320], allow for direct assessment of miRNA-mRNA interactions in samples of interest. These experimental methods can potentially identify context-specific, non-canonical, and non-3' UTR binding interactions between miRNAs and mRNAs. Indeed, miR-34a, which is up-regulated by both 6 and 16 week HFDs in our data, can regulate the levels of RXRα via a coding region interaction with this gene [332]. Analysis of AGO CLIP-Seq data with sophisticated biophysically detailed models has revealed substantial numbers of non-canonical binding sites for miRNAs on target mRNAs [346]. However, more recent analysis of such non-canonical interactions suggests that mRNAs possessing these types of sites are no more repressed than mRNAs with no sites at all [347]. Thus, complementation of computational methods for target

site prediction with direct experimental approaches may improve predictions drawn from the methods applied here, though our analyses provided a wealth of new hypotheses that can be validated with direct experiments.

Additional modeling strategies and complementary dataset integration may also improve our understanding of miRNA-mediated regulation during hepatic insulin resistance. Biochemical considerations, including miRNA and target mRNA abundances and duplex binding strengths, likely play a major role in determining the influence of miRNA regulatory networks. Detailed mathematical modeling of target site abundance effects suggests that higher concentrations of targets may titrate the availability of miRNAs for effective repression [348]. Additional mathematical models have also provided insights into how miRNA regulation fine tunes downstream protein expression levels [349, 350]. In our modeling efforts, we weighted miRNAs by their level of change between diets (i.e. $\log_2$ fold-changes) but did not explicitly consider abundance effects. While TargetScan provides continuous scores reflecting the predicted levels of transcriptional repression expected from a given miRNA-target site interaction, additional biochemical considerations may improve modeling of these regulatory interactions. Additionally, we did not incorporate proteomic-level data in our SAMNet models. Such data added as extra regulatory layers could enhance our understanding of these regulatory networks by incorporating evidence of miRNA-induced effects on transcription factor protein levels directly, as well as effects on down-stream protein levels due to altered transcription via these factors. Unfortunately, our proteomic data collected from these samples did not quantify many transcription factor proteins due to their low abundances, rendering the first proposed regulatory layer difficult to implement. Future studies could specifically quantify the levels of target transcription factor protein levels to further constrain models and enhance predictions.

In summation, we collected small RNA-Seq data to quantify miRNA expression changes in response to HFD in the liver. We integrated these data with mRNA-Seq and epigenetic data using an enrichment method and a network modeling algorithm to identify critical roles for miRNAs altered by HFD. These methods provided a wealth of knowledge that can be used to drive further study of miRNA regulatory networks in the context of hepatic insulin resistance.

5.4. MATERIALS AND METHODS

5.4.1. Animals

We obtained male C57BL/6J mice (stock number 000664) from the Jackson Laboratories. All mice were housed in a specific pathogen-free facility accredited by the American Association for Laboratory Animal Care. We fed the mice a standard chow diet (Prolab Isopro RMH 3000, Purina) for 24 weeks or a high-fat diet (S3282, Bioserve) starting at 8 weeks for 16 weeks HFD or at 12 weeks for 6 weeks HFD. We euthanized all mice at 24 weeks after an overnight fast and froze the livers prior to removal using clamps cooled in liquid nitrogen. The frozen livers were then pulverized into a powder using a CryoPREP impactor (Covaris). We prepared aliquots of pulverized liver for all samples for subsequent analyses. All experiments were carried out in accordance with guidelines for the use of laboratory animals and were approved by the Institutional Animal Care and Use Committees (IACUC) of the University of Massachusetts Medical School.

5.4.2. Small RNA-Seq data collection and analysis

We extracted total RNA from the pulverized frozen livers of mice (three per condition) using the miRNeasy Mini kit (Qiagen). We prepared small RNA-Seq libraries using the NEBNext® Small RNA Library Prep Set for Illumina with 1 µg of total RNA. These libraries were multiplexed and single-end sequenced for 40 base-pairs on an Illumina Hi-Seq 2000 machine. On average, we obtained ~10 million raw sequencing reads for each individual library.

We built a custom analysis pipeline to quantify mature miRNA read counts in samples. First, raw reads were trimmed at their 3' ends to remove excess adapter sequences and contamination reads (e.g. adapter dimers or too long [28 bp] or short [16 bp] reads) were discarded using the clipper tool in the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). We then aligned the clipped and filtered reads to known mouse miRNA hairpin sequences obtained from version 18 of miRBase [351] using the short-read alignment tool Bowtie (version 0.12.7) [352] with the parameters "`--solexa1.3-quals -S -v 1 -q -a --best --strata --norc`". Reads that

aligned within a +/- 3 base pair offset to the annotated -5p or -3p mature miRNA positions within these hairpin sequences were retained and added to the total read count for each mature miRNA. Ambiguous alignments, i.e. those that mapped to >1 mature miRNAs, were either discarded if they mapped to mature miRNA sequences that differ or retained if they mapped to different mature miRNAs that have the exact same sequence (e.g. miR-3107-5p and miR-486-5p). For the latter case, we combined the records for these miRNAs into a single combined species. We created a table of mature miRNA counts for each sample and used DESeq2 (version 1.0.18) [227] to find differentially expressed miRNAs between conditions. We considered a miRNA to be differentially expressed if it possessed an absolute $\log_2$ fold-change between conditions $\geq$ 0.32, an FDR-adjusted p-value (q-value) $\leq$ 0.05, and possessed at least 10 raw read counts in at least one tested condition.

5.4.3. mRNA-Seq and 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 [122] (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) [225]. We restricted TopHat to only align to known transcript splice junctions. We used the Bioconductor package conditional quantile normalization (CQN, version 1.6.0) [226] to remove systematic biases due to GC-content and gene length coverage and used DESeq2 (version 1.0.18) [227] to perform differential expression analyses. We considered a gene to be differentially expressed if it possessed an absolute $\log_2$ fold-change between conditions $\geq$ 0.5, an FDR-adjusted p-value (q-value) $\leq$ 0.05, and was expressed in at least one tested condition (i.e. $\geq$ 0.1 FPKM).

5.4.4. Histone modifications and determination of transcription factor gene targets

These methods are described in detail in ***Chapter 2*** (sections 2.4.5 and 2.4.8) of this document. We used the transcription affinity scores calculated for each factor motif against each differentially expressed gene between CD and 16 week HFD as the weights connecting transcription factors to genes in the SAMNet modeling formulation described below.

5.4.5. miRNA-mRNA target predictions

We generated custom miRNA-gene target predictions by running Perl scripts obtained from TargetScan (version 6.0) [301, 319]. We extracted miRBase (version 18) mature miRNA sequences and mm9 3' UTR sequences from RefSeq gene models (from January 25, 2012) for use with the `targetscan60.pl` (to predict targets) and `targetscan_60_context_scores.pl` (to compute context+ scores) scripts. We retained predicted miRNA-gene target pairs possessing a total context+ score < -0.1. By TargetScan convention, the more negative the score, the more the interaction is expected to influence target repression.

5.4.6. miRNA-mRNA target enrichment analysis

We queried the sets of differentially expressed mRNAs in our mRNA-Seq datasets for predicted targets of each differential miRNA using our custom TargetScan miRNA-mRNA target predictions (context+ score < -0.1). We only considered miRNA-mRNA target predictions in which the two species in question were anti-correlated in terms of expression $\log_2$ fold-changes in HFD versus CD. We used the hypergeometric distribution to compute enrichment statistics for overrepresentation of predicted targets for each miRNA:

$$MirPV = \sum_{i=\text{numTargets}}^{\min(K,N)} \frac{\binom{K}{i}\binom{M-K}{N-i}}{\binom{M}{N}}$$

where *MirPV* is the upper-tail hypergeometric p-value for the current miRNA, K is the total number of differential mRNAs that are anti-correlated with respect to the current miRNA, N is

the total number of predicted mRNA targets for the current miRNA, *numTargets* is the number of identified anti-correlated predicted mRNA targets (i.e. the overlap of *K* and *N*), and *M* is the total number of expressed genes. We then corrected these p-values for multiple hypotheses testing using the Benjamini-Hochberg FDR procedure [257] and reported miRNA enrichment rankings as an ascending list of corrected p-values.

5.4.7. SAMNet modeling

We used the SAMNet algorithm [162] to jointly model our small RNA-Seq, mRNA-Seq, and epigenetic data, adapting the structural layout of Gosline *et al.* (2016) [324]. In this formulation (see **Figure 5-3**) nodes represent differential miRNAs, transcription factors, active epigenetic regions, and mRNAs, while edges were assigned weights for each regulatory layer using 1) miRNA expression $\log_2$ fold-changes, 2) predicted miRNA-mRNA target scores from TargetScan (considering only anti-correlated relationships), 3) mRNA expression changes in genes encoding transcription factors, 4) predicted affinity scores for transcriptional regulators near differentially expressed genes derived from our epigenetic data set analyses, and 5) mRNA expression $\log_2$ fold-changes.

SAMNet drives "flow" through this constructed graph, which consists of all the possible interactions amongst the data, beginning at an artificial source node, flowing through the data layers, and collecting at an artificial terminal sink. The algorithm finds a compact network representation by using constrained optimization to find the best path from source to sink that balances the inclusion of many data nodes at the expense of many high cost edges with including only high confidence edges at the expense of excluding many data nodes. A tuning parameter γ is utilized to balance these competing objectives; generally, larger γ values produce larger networks. We generated networks using γ = [12, 13, 14, 15, 16, 17, 18]. For the final network shown in **Figure 5-4**, we used a SAMNet solution run with γ = 16 as this model incorporated many of the differential miRNAs and solutions run with larger values differed only minimally from this run. For presentation, we only show the miRNA-transcription factor regulatory layer for simplicity, removing the artificial, epigenetic, and mRNA nodes.

CHAPTER 6

GENERAL CONCLUSION

6.1. SUMMARY, DISCUSSION, AND IMPLICATIONS

This thesis primarily presented applications of quantitative, multi-omic systems biology approaches to the study of obesity-induced hepatic insulin resistance. This work is novel in terms of 1) the breadth of the omic levels profiled, encompassing the transcriptome (both mRNA and miRNA expression), epigenome, global proteome, and metabolome, and 2) the level of simultaneous multi-omic integration and computational modeling applied to such data in these contexts. Throughout, I described methods that are readily scalable and applicable to the holistic study of diverse biological phenomena. I anticipate that these methods and the results gleaned from their application will be useful to a wide variety of experimental and computational scientists.

Chapter 2 described the major thrusts of my graduate work. It is here that I presented analyses of the hepatic transcriptomes, epigenomes, proteomes, and metabolomes of chow diet and long-term high-fat diet-fed mice, along with subsequent multi-omic integration of these datasets with the PCSF network modeling algorithm. I uncovered changes induced by HFD at each omic regulatory level individually and, importantly, I found distinct differences in terms of the biological processes implicated by each dataset in isolation. For instance, I showed that mRNA and protein changes induced by HFD are only modestly correlated; thus, the use of mRNA information as proxies for protein-level information is generally unreliable without knowledge of additional biochemical parameters (e.g. translation and degradation rates). These results are consistent with observations made by others in this realm and elsewhere [106, 177]. Thus, it is critical that we design studies that consider multiple types of information to truly gain a holistic view of disease.

Integration of these omic levels with computational network modeling uncovered a highly interconnected web of biological processes and pathways affected by obesity. Novel aspects of this modeling were 1) the inclusion of protein-metabolite interactions derived from the human metabolome database [117] with protein-protein information (from iRefIndex [110]), 2) the use of negative prizes on interactome molecules to enhance the specificity of network solutions by avoiding "frequent fliers" in network results, and 3) the implementation of new criteria to aid model selection. This latter task is especially difficult because no true gold-standard against which we can validate network results exists. In work not presented in the previous chapters of this thesis, I found that members of well-established canonical signaling pathways possess many interactions beyond those conveniently included in such interaction models, highlighting the level of cross-talk between molecules and pathways in true biological systems. In addition, and in contrast to many other systems biology studies in this realm that applied some form of network modeling, I did not include mRNA expression data directly into my models due to the weak mRNA-protein correlations I found. Rather, this data was used, along with epigenetic data of histone modifications, to infer transcriptional regulatory proteins that likely influence downstream gene expression changes induced by HFD. Transcriptional datasets are still highly informative, however, as they contain latent information of upstream regulatory actions. Here, I leveraged transcriptional datasets for these purposes.

Clustering of our final PCSF model revealed twenty sub-networks with unique enrichments for various biological processes. These results critically stress that complications associated with obesity-induced insulin resistance and type 2 diabetes are not restricted to well-studied pathways and processes alone. Indeed, I reviewed recent work that has proposed novel mechanisms of hepatic insulin resistance that complicate traditional views of such processes [45-48]. I found enrichments in pathways with well-established relationships to hepatic insulin resistance, including glucose metabolism, amino acid metabolism, fatty acid metabolism, and transcriptional regulation. More interestingly, I found alterations to a number of biological processes and components that are typically not considered in these contexts, including cell-cell communication, the extracellular matrix, bile acid metabolism, and apoptosis. In addition, we found a sub-network enriched in unfolded protein response (UPR) molecules; Wu *et al.* (2014) recently found evidence for dysregulation of the UPR in these contexts [177]. We tested and

validated a number of these predictions with follow-up imaging studies of CD and HFD mouse hepatic tissue and found dysregulation of general hepatic architecture, tight junctions, and bile acid metabolism, along with enhanced hepatocyte apoptosis in HFD livers. Thus, these results suggest directions for future study of the liver's role in type 2 diabetes pathogenesis.

Chapter 3 focused more specifically on transcriptional regulation in the liver. We compared mice fed CD, short-term (6 week) HFD, and long-term (16 week) HFD treated without and with the type 2 diabetes drug metformin. My transcriptional analysis found that HFD progressively induced transcriptional dysregulation in the liver as time on this diet increased. Metformin only had modest effects on hepatic transcription, though some of the metformin-sensitive genes identified (e.g. *Cebpb*) indeed play a role in regulating hepatic glucose handling and may be critical to this drug's mechanisms of action *in vivo* [242].

We additionally performed temporal profiling of hepatic transcription following stimulation with intraperitoneal insulin in CD and 16 week HFD mice. As may be expected, insulin induced a robust transcriptional response in the liver, with >800 genes changing in expression following stimulation. This effect was almost completely abolished in HFD livers, with only 29 genes changing significantly in response to insulin in both diets. However, we did find a set of 137 genes that significantly responded to insulin in HFD livers alone. Among these is the regulator of G-protein signaling 4 (*Rgs4*) gene, which has previously been studied extensively in the context of neurological processes and disorders [248-252]. *Rgs4* expression was uniquely induced by insulin in HFD livers according to our temporal transcriptomic data; we validated these findings in mice and established that *Rgs4* is indeed expressed and sensitive to hormone (insulin) and cytokine (TNFα) treatment in primary mouse hepatocytes. We additionally tested this response in mice lacking the liver insulin receptor and found that this effect of insulin on *Rgs4* expression in HFD mice depends on this gene. These results together expand upon the notion that hepatic insulin signaling is actually intact in the insulin resistant state. This phenomena has been described as "selective insulin resistance," whereby only the branch of insulin signaling that controls hepatic glucose production is dysregulated, while the lipogenic effects of insulin remain intact [40]. Others have proposed that insulin signaling is nearly fully intact during insulin

resistance and that mechanisms dependent on signals from other tissues (e.g. adipose) hamper insulin's ability to control glucose output [46-48].

We further characterized the role of RGS4 in the liver by feeding mice lacking this gene a HFD. We observed a significant effect of *Rgs4* deletion on hepatic insulin sensitivity, whereby mice lacking this gene were more insulin resistant compared to wild-type mice during an insulin tolerance test. We proposed that this effect is mediated in part by RGS4's ability to inhibit G-protein signaling and subsequent activation of PKC [247]. PKC is known to inhibit the kinase functions of the insulin receptor, and knock-down of this target protects rats from hepatic insulin resistance [38]. Indeed, we found that PKC activity is elevated in 16 week HFD-fed mice. Thus, we identified a potentially novel mechanism that promotes hepatic insulin sensitivity in the face of enhanced liver fatty acid content. Our group is currently performing additional analyses to further characterize the role of RGS4 in the liver. Additional findings may reveal new avenues for therapeutic intervention based around this molecule's action.

Chapter 4 described analyses of hepatic transcription and epigenetics in mice fed a calorie-restricted diet, in addition to mice fed the CD and long-term HFDs discussed above. Caloric restriction has been shown to extend lifespan, improve insulin sensitivity, and delay the onset of age-related diseases, including diabetes [262, 263]. Mice fed a CR diet lost weight compared to CD controls and are generally considered healthy, especially in comparison to HFD mice. I compared transcriptional changes induced by HFD and CR versus CD and found that both diets induce extensive changes in gene expression. Most interestingly, I uncovered a significant sub-set of genes modulated by both HFD and CR that change in the same direction compared to CD. This included genes that promote fatty acid, lipid, and cholesterol synthesis. Up-regulation of such genes is a well-characterized consequence of HFD [353]. Interestingly, CR mice also enhance fatty acid synthesis to promote subsequent energy production via β-oxidation as they intake less overall energy from food [281].

In addition to transcriptional data, we collected chromatin accessibility data by DNase-Seq to profile active regulatory regions throughout the genomes of CD, HFD, and CR mice. I used bioinformatics sequence analysis of the discovered accessible regions near genes modulated by

the diets to infer regulators that may play a role in controlling differential transcription. Somewhat surprisingly, I found nearly identical transcription factor enrichments near all the gene sets tested. These results imply that a relatively small set of factors are capable of responding to various dietary challenges to alter gene expression. We particularly focused on enrichments for nuclear hormone receptor family members by performing follow-up ChIP-Seq experiments in HFD and CR mice for PPARα and RXRα. We indeed found that these two factors bind extensively throughout the genomes of these livers. In particular, we noted extensive binding of PPARα near genes involved in glucose metabolism, a number of which whose expression levels were modulated by diet. We further tested the role of PPARα *in vivo* and *in vitro* by treating mice and primary mouse hepatocytes with the PPARα activator fenofibrate. *In vitro* results demonstrated that PPARα plays a role in enhancing anaerobic glycolysis. Our *in vivo* results validated several novel predicted gene targets of PPARα involved in glucose metabolism. Thus, this multi-omic study identified new roles for the PPARα transcription factor in the control of glucose metabolism in the liver.

Finally, in ***Chapter 5***, I examined HFD-induced changes in mouse hepatic miRNA expression profiles. Similar to results observed for mRNAs in ***Chapter 3***, HFD feeding progressively modulated miRNA expression patterns. To identify differential miRNAs that may play important functional roles following HFD, I devised an enrichment scheme whereby miRNAs were ranked by the prevalence of target mRNAs present in the pool of HFD-sensitive genes. I used the TargetScan algorithm to computationally establish gene targets of these miRNAs [319]. Among the high-ranking miRNAs established by this approach were miR-34a, miR-152, and miR-378, all of which are known to regulate targets involved in processes related to obesity and insulin resistance [306, 332-335]. Importantly, this approach also identified miRNAs with less established relevance in this context, including miR-1839-5p, miR-149-5p, miR-455-5p, and miR-532-5p. These miRNAs may indeed perform critical functions in the liver that either promote or relieve complications associated with obesity.

We also employed a flow-based network modeling method called simultaneous analysis of multiple networks (SAMNet) [162, 324] to directly interrogate miRNA-transcription factor regulatory networks. This approach integrated miRNA expression, miRNA-mRNA target

predictions, mRNA expression, and transcription factor affinity information inferred from histone modification and motif data. This analysis identified a number of potentially relevant miRNA actions. In fact, the majority of the interactions predicted by this model, to the best of my knowledge, have not been investigated in this context by prior studies. Thus, this work presented methods by which miRNAs can be functionally analyzed in context with additional omic measurements. I particularly assessed such miRNA-mediated regulatory activities in the context of obesity-induced hepatic insulin resistance and suggested avenues for future study.

6.2. LIMITATIONS AND FUTURE PERSPECTIVES

As is the case with all scientific endeavors, limitations in terms of the chosen model system and methods used must be considered. To start, all of these results and insights derive from studies in mice. It is probable that metabolic differences between mice and humans contribute to variable responses to HFD between these species. Indeed, in the context of cancer, evidence for differences in tumorigenesis between humans and mice has been shown [354]. However, given the limited availability of human tissue samples, the amount of material required to collect the breadth of data described here, the ability to stringently control genetic and environmental factors, and the similarities between mice and humans in terms of observed pathologies on the road to metabolic syndrome following obesity, these studies present crucial insights into general hepatic complications that promote the insulin resistant state. Future human studies are necessary to truly assess the translational aspects of our findings.

In these studies, we used whole-liver tissue from CD and HFD (and CR) mice as our physical material. An advantage of this is that we captured as closely as possible the true *in vivo* state of the system. However, differences do still manifest between mice raised in the same environment, which can introduce noise into the system that may mask important molecular changes. Also, a multitude of liver cell types beyond hepatocytes (which comprise ~90% of the liver) exist, and stressors like HFD activate dormant cell types that contribute to molecular changes. Such cell-type specific contributors are difficult to ascertain. While computational methods attempting to de-convolve cell type contributions in complex tissues exist [355, 356], they typically require *a priori* data from purified reference populations and generally use linear methods to estimate

proportional contributions. Such linear assumptions likely do not hold true in all cases as positive and negative feedback mechanisms induce synergistic and/or antagonistic effects on molecules.

The goal of omic data collection is to comprehensively quantify a given layer of biological regulation in as unbiased a fashion as possible. RNA-Seq has been shown to be a highly reproducible tool for profiling the transcriptome [80], though biases due to sequence content (e.g. GC content) can influence observed expression patterns [226]. In this work, I applied computational methods to limit biases due to GC content and gene length effects when calling differentially expressed genes. This was critical because I observed a significant correlation between sequence GC content and estimated $\log_2$ fold-changes between conditions when not correcting for these influences, perhaps due to batch effects. Additionally, sequence effects hampered some early epigenetic data we collected, whereby we observed severe read coverage depletion in annotated CpG islands across the genome. Thus, these are critical considerations that must be made when collecting any high-throughput sequencing datasets. In some cases these effects can be corrected with statistical modeling, while in severe cases experiments may need to be repeated to assure proper quality and downstream interpretation.

Our proteomics data quantified >50,000 unique peptides that mapped to >6,000 unique proteins. While this is indeed a rich dataset, there are still many expressed liver proteins that were not quantified, likely due to low abundance. In particular, we either did not observe or only quantified a small number of peptides for many transcriptional regulators that are known to be expressed in the liver, including FOXO1, HNF1A, HNF1B, and PPARα. Missing data is also common in such proteomics datasets, requiring either imputation methods to fill in gaps or outright removal of poorly covered peptides. In addition, technical noise combined with biological noise may have masked some true signals. To this, we additionally collected phospho serine and threonine data from CD and long-term HFD samples and, although we quantified ~9,000 modified peptides, we were unable to reliably detect any statistical changes due, potentially, to measurement noise. I noted that the changes between CD and HFD livers that we did observe are consistent with a number of other studies that performed both targeted and untargeted proteomics in this context, but we are likely missing some important measurements that may influence the insulin resistant state. Similarly, we only quantified ~400 small molecules

in our metabolomics data. As proteomic and metabolomic data collection methods improve and become more sensitive [357, 358], additional insights may be revealed from deeper analyses of these omic levels.

My network modeling efforts indeed revealed unique insights into hepatic insulin resistance; however, mechanistic interpretation of the included interactions is still limited due to incomplete information. For instance, a number of interactions derive from high-throughput experiments, which provide little additional information (e.g. directionality or activation/inhibition) beyond simple binary interaction calls. More detailed mechanistic modeling methods require a variety of additional parameters to simulate network behavior, including molecular concentrations, sub-cellular locations, binding strengths (i.e. K_D), catalytic rate constants, etc. In work performed by Minyi Lee, an MIT undergraduate student who worked with me as part of the MIT UROP program, we attempted to apply mechanistic modeling principles to small sub-networks that would arise from network modeling of omic data. We started with a model of IL-6 signaling in hepatocytes [359], whereby we extracted the protein-protein interaction information of this pathway from the interactome and attempted to fit this model to "data" derived from the well-parameterized version of this model. This procedure generally did a good job of matching the characteristics of the "true" model. Minyi also applied this methodology to a sub-network of EGFR signaling generated from network modeling of time-series phospho-proteomic data. Again, parameter fitting produced a model that fairly accurately matched the kinetics and trajectories of the true measurements. Additional simulations with perturbations are needed to accurately assess overall model performance. Thus, a goal for the future in general is to produce more detailed interactomes that enable such mechanistic modeling efforts. Indeed, considerable effort towards inferring causality in networks is an active area of computational systems biology research, though methods thus far have mostly been applied to small, well-defined pathways that may not scale to larger interaction networks [360]. In the future, network modeling methods, including the PCSF, could be run on directed interactomes where edge scores reflect some form of biophysical quantity (e.g. K_D) and where node prizes are assigned weights according to their concentrations in specific cell types or tissues (in addition to correlation with disease). Indeed, Hein *et al.* (2015) applied a method called quantitative bacterial artificial chromosome green fluorescent protein interactomics (QUBIC [361]) in HeLa cells to identify specific interactions of

>1,000 bait proteins, measuring interaction stoichiometries and cellular abundances [362]. This allows for classification of stable versus transient interactions and provides additional data dimensions from which physical interactions can be characterized. This type of data can enhance subsequent down-stream interpretation, modeling, and hypothesis testing of disease networks.

Here we profiled a wide array of biomolecules using several omic methods. Still, additional omic layers are quantifiable and likely involved in the pathology of type 2 diabetes. For instance, obesity is associated with changes in the gut microbiome and fecal transplantation from lean donors into patients experiencing metabolic syndrome has been shown to improve insulin sensitivity [120, 363]. More mechanistically, changes to mouse gut microbiota affect FXR in the ileum and modulate bile acid metabolism in the liver [121]. Additional data types, including methylome data and cytokine expression profiles, may enhance systems analysis of type 2 diabetes.

6.3. CLOSING REMARKS

Obesity-induced hepatic insulin resistance is a highly-complex condition that involves coordinated dysregulation of many molecular entities across a spectrum of biological regulatory levels. We strove towards a holistic understanding of this metabolic condition by applying a systems biology approach to the study of diet-induced obesity. The results presented in this thesis provide new insights into how this condition manifests molecularly and will hopefully drive future exploration of approaches to treat diseases like type 2 diabetes.

APPENDIX A

NETWORK-BASED INTERPRETATION OF DIVERSE HIGH-THROUGHPUT DATASETS THROUGH THE OMICS INTEGRATOR SOFTWARE PACKAGE

This work has been published as:

Nurcan Tuncbag*, Sara J.C. Gosline*, Amanda Kedaigle, **Anthony R. Soltis**, Anthony Gitter, and Ernest Fraenkel, "Network-based interpretation of diverse high-throughput datasets through the Omics Integrator software package," *PLoS Computational Biology*, 12(4): e1004879. doi: 10.1371/journal.pcbi.1004879. April 20, 2016.

*Denotes equal contribution.

Author contributions: Conceived and designed the experiments: SJCG NT AG AK EF. Performed the experiments: SJCG NT. Analyzed the data: SJCG NT. Contributed reagents/materials/analysis tools: SJCG NT ARS AK. Wrote the paper: SJCG NT AG AK ARS EF. Designed, developed scripts: NT SJCG ARS AK.

Acknowledgments: We thank all the beta testers for running the software, in particular Jonathon Gulliver for his evaluation of the test procedures.

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package

Nurcan Tuncbag[¤a☯], Sara J. C. Gosline[¤b☯], Amanda Kedaigle, Anthony R. Soltis, Anthony Gitter[¤c], Ernest Fraenkel*

Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America

☯ These authors contributed equally to this work.
¤a Current address: Graduate School of Informatics, Department of Health Informatics, Middle East Technical University, Ankara, Turkey
¤b Current address: Sage Bionetworks, Seattle, Washington, United States of America
¤c Current address: University of Wisconsin-Madison and Morgridge Institute for Research, Madison, Wisconsin, United States of America
* fraenkel-admin@mit.edu

Citation: Tuncbag N, Gosline SJC, Kedaigle A, Soltis AR, Gitter A, Fraenkel E (2016) Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package. PLoS Comput Biol 12(4): e1004879. doi:10.1371/journal.pcbi.1004879

Editor: Andreas Prlic, UCSD, UNITED STATES

Received: July 17, 2015

Accepted: March 23, 2016

Published: April 20, 2016

Data Availability Statement: The expression and phosphoproteomic data have been retrieved from the work by Thomson et al., 2011 (doi: 10.1007/s10585-010-9367-3). The epigenetic data have been downloaded from ENCODE. The direct links to these data are ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDnase/wgEncodeUwDnaseA549PkRep1.narrowPeak.gz and ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDnase/wgEncodeUwDnaseA549PkRep2.narrowPeak.gz

Funding: This work was supported by NIH grants U54CA112967, U01CA184898 U54NS091046,

Abstract

High-throughput, 'omic' methods provide sensitive measures of biological responses to perturbations. However, inherent biases in high-throughput assays make it difficult to interpret experiments in which more than one type of data is collected. In this work, we introduce Omics Integrator, a software package that takes a variety of 'omic' data as input and identifies putative underlying molecular pathways. The approach applies advanced network optimization algorithms to a network of thousands of molecular interactions to find high-confidence, interpretable subnetworks that best explain the data. These subnetworks connect changes observed in gene expression, protein abundance or other global assays to proteins that may not have been measured in the screens due to inherent bias or noise in measurement. This approach reveals unannotated molecular pathways that would not be detectable by searching pathway databases. Omics Integrator also provides an elegant framework to incorporate not only positive data, but also negative evidence. Incorporating negative evidence allows Omics Integrator to avoid unexpressed genes and avoid being biased toward highly-studied hub proteins, except when they are strongly implicated by the data. The software is comprised of two individual tools, Garnet and Forest, that can be run together or independently to allow a user to perform advanced integration of multiple types of high-throughput data as well as create condition-specific subnetworks of protein interactions that best connect the observed changes in various datasets. It is available at http://fraenkel.mit.edu/omicsintegrator and on GitHub at https://github.com/fraenkel-lab/OmicsIntegrator.

This is a *PLOS Computational Biology* Software paper.

R01GM089903 and used computing resources supported by the National Science Foundation under Award No. DBI-0821391. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

High-throughput technologies are now able to provide comprehensive and quantitative measurements of molecular changes in response to perturbations or disease. Measurements of the transcriptome, epigenome, proteome, etc. serve to complete the opaque picture of the many active pathways and processes in cellular systems. However, no single dataset fully captures all aspects of cellular activity in a given experimental setting. For example, transcriptional datasets allow us to see which genes are up- or down-regulated relative to a control state, but do not provide information about post-translational modifications that are critical for signaling. In addition, high-throughput datasets often contain many 'hits' (i.e. species that change significantly between conditions in a given omic dataset) that either lie in unexpected pathways [1,2] or fail to map to any existing canonical pathways [1–5]. Thus, in order to discover novel biological processes associated with specific perturbations or disease [6–8] we need to consider data from complementary high-throughput datasets jointly.

Network modeling approaches allow us to overcome such limitations because they can combine multiple types of data without requiring prior pathway information. These approaches can either use user-generated or publicly available data, such as protein-protein interactions and epigenetic data, to find either direct or indirect (i.e. via unobserved molecules) connections between experimental hits. Critical to these approaches are protein-protein interaction databases, which collate data from multiple experimental platforms and cell types to provide networks of experimentally detected interactions [9–12]. In addition, recently generated protein-metabolite and protein-small molecule interaction networks, including HMDB [13], DrugBank [14] and STITCH [15], allow for richer assessment of molecular interaction types that occur in cellular systems. These collections of physical molecular interactions, or interactomes, enable researchers to apply network modeling approaches to a wide variety of data.

Measurements of transcriptional changes in response to perturbation or disease are a commonly generated omic data type. However, the proper approach for including transcriptional measurements in networks requires some thought. Since these data do not directly measure protein abundance or activity, it is misleading to map them directly to their corresponding proteins in the interactome. Instead, Omics Integrator combines such transcriptional data with epigenetic data to identify putative changes in the activity of DNA binding proteins that influence transcriptional changes [16].

This wealth of interaction data gives rise to new challenges. The published interactions between proteins, DNA, and small molecules comprise a network of millions of connections that is a 'hairball,' or a network that is too dense to interpret [5]. There are numerous individual tools that are now available to analyze these networks, each with different capabilities and intended application areas [1,4,17–25] (see Fig 1). Many network optimization methods that aim to reduce hairball interactomes to higher confidence subnetworks exist; however, many of these have limitations that inhibit their general applicability, such as requiring predefined source and target sets [1,4,20,26–29], which is not applicable in cases where omic data do not fit a 'source-target' framework. There are also methods that map mRNA expression datasets to protein interaction networks (MATISSE [30]), methods that identify transcription factor binding sites from epigenetic data (Centipede [22] and PIQ [24]), and methods that relate chromatin features and DNA-binding motifs to gene expression via multivariate/univariate regression (REDUCE [23] and MOTIF REGRESSOR [18]) or support vector regression [31]. These methods fall into two general classes: methods that attempt to reconstruct signaling pathways or interaction networks from data hits or methods that focus specifically on transcriptional regulatory networks. These two classes of tools are both essential to fully integrate diverse types of high-throughput data.

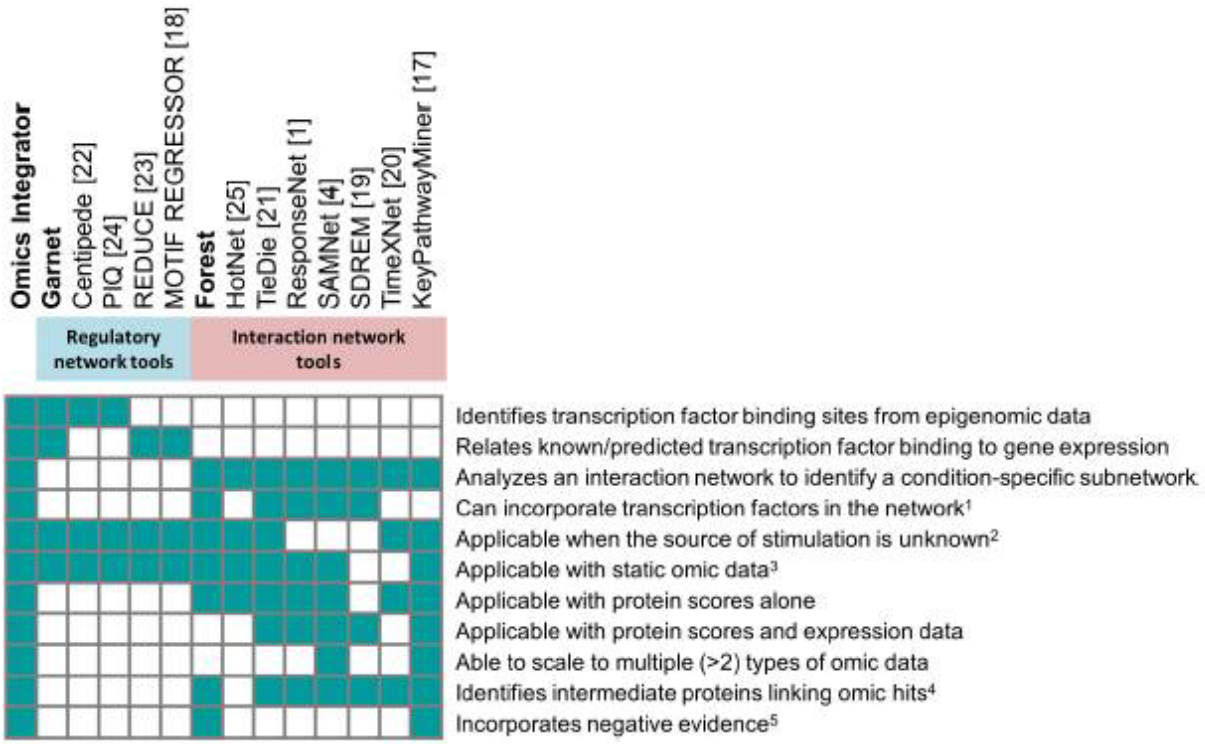


Fig 1. Summary of features differentiating Omics Integrator from existing tools and which features are available when Garnet and Forest are used individually. [1] Some network algorithms model TFs by including protein-DNA interactions in the network or generating TF scores for the protein nodes. [2] Some network algorithms optimize the transmission of information from source nodes to target nodes and require the sources to be identified in advance. [3] Time series analysis algorithms require omic data from three or more time points. [4] Intermediate proteins, like the Steiner nodes predicted by Forest, are not assigned condition-specific scores but are important for connecting other scored nodes in the subnetwork. [5] Negative evidence discourages network algorithms from selecting particular nodes due to prior knowledge or a bias, such as node degree.

doi:10.1371/journal.pcbi.1004879.g001

The tools currently available for omic data integration provide only a subset of the features provided by Omics Integrator (Fig 1). Furthermore, there is no existing algorithmic framework that enables the incorporation of weighted negative evidence. We define negative evidence in this context as any data or feature that supports potential exclusion of a species (protein, metabolite, etc.) from a network model solution (e.g. due to lack of expression in the system of interest, etc.). Supporting negative scores can help avoid misleading results in network analysis. KeyPathwayMiner [17] allows a list of negative nodes as input, but these are hard cut-offs. By contrast, Omics Integrator allows for weighted negative scores so the user can balance the prior evidence for excluding a node against the benefits of using it to connect nodes with positive scores.

In summary, we introduce Omics Integrator, a software package that fills a noticeable gap in omics data analysis by providing a unified framework for integrating transcriptomic data together with other omic data using interactome data. Although the individual components, Garnet and Forest, have similarities with existing tools, uniting expression analysis and network analysis in a single package makes it substantially easier to model multiple types of omic data. Omics Integrator expands upon and combines the prize-collecting Steiner forest (PCSF) algorithm [2,5,32] and methods similar to those implemented in previous network algorithms [4,33].

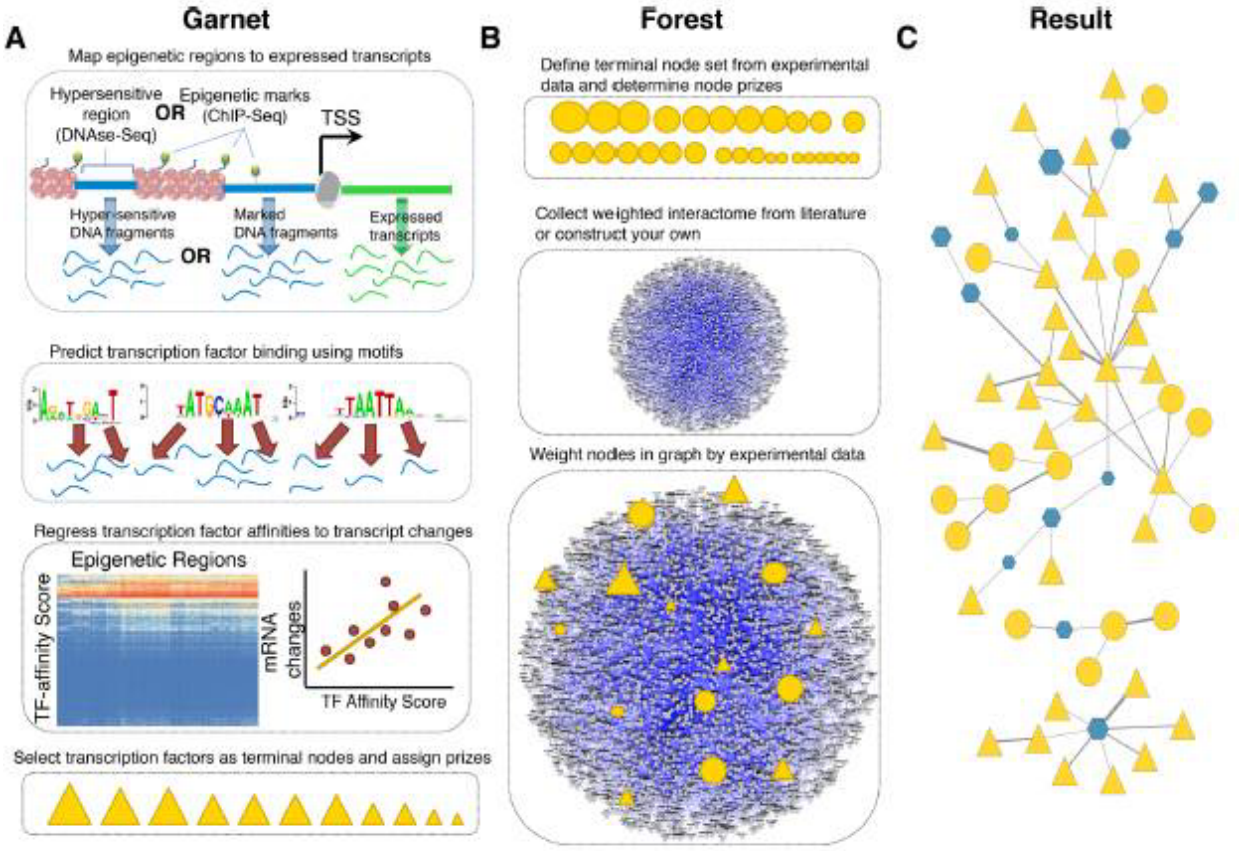


Fig 2. Summary of Omics Integrator **(A)** Garnet identifies transcription factors (triangles) associated with mRNA expression changes by incorporating epigenetic changes nearby expressed genes, scanning those regions for putative transcription factor binding sites and then regressing transcription factor affinity scores against gene expression changes. The result is a set of transcription factor candidates and the relative confidence that they are responsible for the observed expression changes. **(B)** Forest identifies a condition-specific functional sub-network from user data and a confidence-weighted interactome. The network can be composed of protein-protein, protein-metabolite or other interactions. The set of omic hits are composed of the TFs obtained from Garnet (triangles) merged with other types of hits such as differentially expressed proteins, significantly phosphorylated proteins, metabolites, etc. (circles). **(C)** Finally, the confidence-weighted interactome is integrated with the 'omic' hits using the prize-collecting Steiner forest algorithm, where the data is either connected directly or via intermediate nodes, called 'Steiner nodes'.

doi:10.1371/journal.pcbi.1004879.g002

Design and Implementation

The Omics Integrator package (Fig 2) consists of two distinct tools: Garnet and Forest. These tools work together to enable the integration of data derived from measurements of mRNA, proteins, genetic perturbations or metabolites.

Garnet takes chromatin accessibility data (e.g. DNase-Seq, histone modification ChIP-Seq), either generated by the user or acquired from public repositories (e.g. the ENCODE consortium [34], the NIH Roadmap Epigenomics Mapping Consortium [35]), and identifies a set of transcriptional regulators that potentially explain observed gene expression levels or changes between conditions in an experiment of interest (Fig 2A). From the chromatin accessibility data, Garnet scans regions proximal to transcribed genes for transcription factor binding sites. Binding sites are inferred from sequence matches in the underlying DNA to a clustered set [36] of DNA-binding motifs. Garnet then uses a pseudo-thermodynamic metric [37] to compute a

transcription factor affinity (TFA) score in these regions and maps these scores to genes within a fixed window (e.g. 2kb). Garnet estimates transcription factor activity by performing univariate linear regression of the TFA scores against corresponding mRNA expression measurements. Significant regression coefficients indicate candidate transcriptional regulators that can then be used as input to the Forest program or analyzed independently.

The Forest tool identifies a parsimonious interaction network connecting a subset of user-defined omic data hits (Fig 2B). These hits can be the transcription factors selected by Garnet and/or any other type of biological data (e.g. siRNA hits, phosphoproteomic changes, metabolites, etc.). Specifically, Forest solves the prize-collecting Steiner forest problem [5] that takes into account the importance assigned to the omic hits (e.g. by significance level or fold-change between conditions) as well as the probability that each reported interaction is real. Each omic hit and/or Garnet-selected transcription factor is given a positive 'prize', reflecting the confidence in the reliability of the underlying data (e.g. a set of differentially expressed proteins). These user-defined hits are referred to as terminals. When the algorithm includes a terminal in the network, it is rewarded with the prize assigned to that hit, but also has to pay costs for the interactions used to link the data to the rest of the network. By seeking to maximize the collected prizes and minimize the edge costs, the algorithm uncovers a high-confidence set of physical interactions that explain how the omic hits are related. Because the algorithm is not forced to include all omic hits, it can remove those that are poorly connected and give rise to the 'hairball'. However, Forest can at the same time select additional, interactome-derived nodes called 'Steiner nodes' when necessary. Steiner nodes are likely relevant to the biological response in question, but may be missed by the high-throughput assays. Forest generates output files that can be easily viewed using the network visualization software package Cytoscape [38].

Forest includes numerous features to efficiently generate biological networks, while avoiding common pitfalls often encountered when implementing network-based algorithms. For example, many network methods are inherently biased towards using nodes that have been studied more extensively and have more reported interactions [39–41]. Forest can penalize these highly-connected nodes, termed 'hub' nodes here, by assigning to them a prize with a negative value. Without the negative prizes, hub proteins are often selected even when they are not biologically relevant or interesting. With negative prizes, they only appear in the network when other data strongly implicate them. Additionally, Forest includes randomization strategies to ensure that the resulting networks are robust to noise in the biological data. We include in our software package a straightforward example in which phosphoproteomic measurements are integrated with changes in mRNA expression [42] using epigenetic data from a related cell line from ENCODE [34,43]. We also include examples in which Garnet and Forest can be run independently, described in more detail on our website: http://fraenkel.mit.edu/omicsintegrator.

Mapping gene expression changes to proteins

The Garnet algorithm reveals candidate transcriptional regulators that likely influence gene expression levels or changes. We prefer to not use mRNA measurements as proxies for changes in protein levels or activities as the relationship between mRNA and protein levels is complex [44–46]. In addition, changes in protein concentrations are not reliable evidence for changes in the activities of pathways, which are often post-translationally regulated. Once Garnet identifies transcription factors that give rise to the observed mRNA changes, the transcription factors can then be used as input to Forest.

Garnet is a core part of omics integration, enabling gene expression changes to be mapped to transcription factors in the interactome that can then be then analyzed alongside other data

[47]. Garnet builds on the rich epigenetic datasets that have been collected through consortia such as ENCODE by using data from histone modifications or open chromatin experiments to restrict the search space for motif matches to areas more likely to be bound by some regulator. This strategy has been used by this lab [48] and others [49] to reduce spurious motif matches and thus to improve the accuracy of transcription factor binding prediction [22]. Although this type of inference cannot eliminate uncertainty about the binding of a TF to a specific site or gene, the cumulative evidence is predictive of the activity of the TF in a transcriptional response. As Omics Integrator solves the prize-collecting variant of the Steiner problem, it can exclude false positives that are difficult to connect to upstream signaling. Omics Integrator also provides options to filter the candidate TFs from Garnet, including checking whether or not the TF is expressed in the tissue or cell type of interest.

Garnet consists of two steps: (1) computationally predicting transcription factor-DNA interactions from epigenetic data and a set of DNA binding motifs and (2) estimating regulator activities by correlating these predicted transcription factor-DNA interactions with mRNA expression changes in genes neighboring the predicted binding sites.

Epigenetic data processing. Garnet first finds genomic regions that likely harbor transcriptional regulatory proteins by searching enriched open chromatin regions for matches to DNA binding protein motifs. These regions are typically identified using epigenetic data either within a similar cell type or from conserved regulatory regions across cell or tissue types [50]. Garnet first searches a list of chromatin regions, provided by the user or included with Omics Integrator, and assigns the regions to genes with open chromatin sites within a user-defined distance threshold (e.g. 2000 base pairs). Garnet then scans these regions to determine the likelihood that a transcription factor will bind a region using a set of position weight matrices (PWMs).

Transcription factor binding prediction. Each region of DNA associated with an epigenetic signal is given a score representing the probability of binding for each motif matrix using a statistical-mechanics framework, where the number of possible binding sites and their PWM scores in each sequence are combined to create a single transcription factor binding probability for each region. The Boltzmann weighted partition function below is used for each motif and chromatin region and has been described in this context previously [37]. The equation to define transcription factor affinity is:

$$TFA_j = \frac{\sum_i e^{w_j \cdot m_i}}{\beta_j + \sum_i e^{w_j \cdot m_i}} \quad (1)$$

TFA_j estimates the probability of binding for motif j using all scoring windows i in the region. w_j estimates the probability that the motif score is not a false positive, m_i represents the PWM log-likelihood score at the i^{th} window in the region, and β_j estimates the probability that motif score is a false positive. In practice we use w_j and β_j as tuning parameters based on the TRANSFAC MATCH minSUM and minFP score thresholds for each motif. For genes with multiple epigenetic regions within the associated window, Garnet chooses the highest TFA_j value for each motif j across all regions. The result is a matrix representing the affinity score for each gene and each transcription factor binding motif.

Transcription factor selection. Garnet uses linear regression to identify motifs mapped to transcription factors with the strongest relationships to the expression data. This approach is similar to that in previous work [37] and assumes that the better the match of a sequence to a motif (summed over an epigenetic region), the stronger the binding and the greater the effect the regulator has on transcription. We apply least-squares regression to relate the *TFA* scores described above to mRNA expression changes for a particular condition of interest.

Significance is assessed by testing the null hypothesis that the slope of the regression line is 0. Transcription factors with motifs exhibiting statistically significant regression coefficients (p-value<0.05 or any desired threshold) are given a weight of–log(p-value).

Network modeling

Selection of the terminal node set and assignment of prizes. To map experimental data to an interactome of interest (Fig 2B), the user must first identify the most biologically significant hits from each dataset and define them as terminal nodes in the network. Terminal nodes are any entities represented by nodes in the network that the user would like the algorithm to analyze in a larger biological context. Typically, these are molecules that change significantly under a treatment relative to an appropriate control. Prizes, denoted p(v), where v is a vertex (node) in the interactome graph, are assigned to the terminal nodes by the user. These prizes can be, for example, the log fold change of proteins in an experiment or negative log of the significance level describing changes between conditions (p-value or q-value). If there is no quantitative information and only a set of terminal nodes is available, users can assign uniform prizes to each terminal.

As already noted, Forest can use node prizes to incorporate negative evidence about the relevance of a node. A priori, it is impossible to know if such a protein has a high degree because it is truly involved in many interactions or because it has been studied more extensively than other proteins because it is highly-conserved, essential, or highly-expressed [39–41]. To avoid the potential bias introduced by these hub nodes, we created a generalized prize function that assigns negative weights to nodes based on the number of connections they have in the interactome. As a result, hubs are less likely to be selected but can still be used when the data strongly support their inclusion. The function for this negative weighting is:

$$p'(v) = \beta \cdot p(v) - \mu \cdot degree(v) \quad (2)$$

where *degree(v)* is the number of connections of node v in the interactome. The β and μ parameters are scaling factors to adjust the effect of terminal nodes and hub nodes in the final network, respectively. When μ is set to 0, the hub correction is disabled (default behavior). Increasing μ attenuates the hub dominance in the optimal solution. Increasing β promotes more terminal nodes to be included in the optimal solution (Eq 2).

In addition to reducing the influence of hub proteins, negative prizes could be used to reduce the influence of molecules that are poorly expressed in a particular tissue or condition. Similarly, negative prizes could be used to exclude molecules that have been experimentally determined to not be relevant to the process under study. Users can take advantage of this feature by simply adding negative values to the original prize file.

Confidence-weighted interactome and edge costs. Calculating the probability p(e) that an edge e between two proteins reflects a real interaction allows us to avoid false positive edges, which are assumed to be less reproducible and therefore less confident. Forest takes as input a set of edge weights (p(e)) and converts them to costs using the scoring function c(e) = 1 –p(e). Several approaches have been described for deriving these probabilities or other confidence scores [1,10,26,51,52].

Forest problem formulation

The input to Forest is a directed, partially directed, or undirected network $G(V, E, c(e), p'(v))$ of node set V and edge set E, where the function $p'(v)$ assigns a prize to each node $v \in V$ and the function $c(e) > 0$ assigns a cost to each edge $e \in E$. The aim is to find a forest $F(V_F, E_F)$ that

minimizes the objective function:

$$f'(F) = \sum_{v \notin V_F} p'(v) + \sum_{e \in E_F} c(e) + \omega \cdot \kappa \quad (3)$$

where $p'(v)$ is as defined in Eq 2, κ is the number of trees in the forest, and ω is a tuning parameter whose purpose is explained below. Here, an artificial node, or 'dummy' node, v_0 is introduced to the initial network and connected to a subset of nodes N (using the --dummyMode option detailed below). The nodes in N are a subset of all the nodes in the interactome V and are assigned a uniform edge cost ω to the dummy node. Forest constrains F to be a tree–a connected graph without cycles–that is rooted at v_0. The optimization problem is solved with the msgsteiner message-passing algorithm to identify a tree subnetwork [53]. Once the network optimization problem has been solved, the root node (v_0) and all its edges are removed, providing a final forest network that is a collection of one or more sub-trees. These sub-trees conceptually represent parallel biological pathways. Given that the resulting solution is dependent on the particular values of β, ω, and μ, we suggest that the user run the algorithm with different settings to select an optimal solution. We recommend choosing optimal parameters based on two criteria. First, the parameters should maximize fraction of terminal nodes included in the network that are also robust to noise; this robustness can be determined by permutations derived from the --noisyEdges flag. Among the parameter settings that yield a similar fraction of robust terminals, we prefer larger networks. It is often also helpful to check that the selected hidden nodes are enriched for biologically-relevant categories using known pathways or gene sets (e.g. from MSigDB [54] or Gene Ontology [55]).

Up to six PCSF parameters are supplied to Forest in a configuration file. The minimum required parameters are ω, β and D. The parameters ω (controlling the number of trees; Eq 3) and β (controlling the trade-off between including more terminals and using less reliable edges; Eq 2) are as described above. D (the depth parameter) controls the maximum path-length from v_0 to terminal nodes. Suggested values for D range between 5 and 10. The optional parameters are μ, g, and *garnetBeta*. The parameter μ controls the degree-based negative prizes (Eq 2); if not provided, μ is assumed to be zero. The reinforcement parameter, g, affects the convergence of the solution (smaller values produce solutions closer to the optimum, but increase run time) and is set to 1e-3 by default. The *garnetBeta* parameter is used to scale the Garnet output prizes relative to the provided protein prizes. The default for *garnetBeta* is 0.01.

Additional algorithm features

Forest converges on a single, optimal solution. However, it can be useful to perform perturbation analyses to determine the robustness of this network and how it relates to suboptimal solutions. Forest provides three different perturbation strategies. To ascribe confidence in the selected hidden nodes, the user can use the --noisyEdges flag to assign noise to the edges of the interaction network. Hidden nodes that appear often in networks run with noisy edge weights are likely more robust than those that only appear in the optimal solution or a small number of noisy runs. A user can also use --shufflePrizes to identify those hidden nodes that are robust to noise in the prize data. Lastly, the user can also assess specificity of hidden nodes using the --randomTerminals flag that runs the optimization with a random selection of terminals (preserving the degree distribution of the original terminal set). Hidden nodes that occur less frequently in forests run with random prizes are likely to be more specific to the user's problem of interest and therefore more biologically meaningful. The results of these perturbations can either be used to weight nodes and edges in the original network using the 'fraction of optimal networks included' attribute or viewed together using the merged network produced by the algorithm.

As previously mentioned, Forest incorporates a dummy node into the graph when solving the optimization problem. When this dummy node is removed after the optimization, the solution is divided into disjoint subnetworks. The --dummyMode option tells the algorithm which nodes in the interactome should be connected initially to the dummy node. The default option (--dummyMode terminal) connects the dummy node to all of the input terminals, guaranteeing that each sub-network in the optimal solution is rooted by a terminal node. The option --dummyMode filename allows the user to explicitly specify which nodes to connect to the dummy node. For example, in our previous work, we used this option to identify signaling pathways originating at cell-surface receptors [5]. There are additional values for the --dummyMode option, recommended for advanced users, as described in Table 1.

Forest can also automatically perform *in silico* knock-out experiments, i.e., identifying new solutions when a specific node removed from the interactome. These can be valuable for determining the robustness of the solution and the significance of individual nodes [5]. To use this option, the user specifies --knockout and provides node names specifying node(s) the user would like to 'knock out', i.e. TP53 or TP53 EGFR.

There are other Forest options in addition to those explained here. The full list is provided in Table 1. The step-by-step procedure to run Omics Integrator and troubleshooting guidance are provided in the **Supplementary Material**. A flowchart showing how to run Omics Integrator is depicted in Fig 3.

Results

Omic data integration and network reconstruction in lung carcinoma

To showcase the utility of Omics Integrator, we analyzed several types of omic data from lung carcinoma cells. We collected previously published data [42] from H358 cells, a model of lung cancer, that were stimulated with TGF-β. Measured gene expression changes were used as input into Garnet together with DNase I hypersensitive regions from A549 cells, a related lung carcinoma-derived cell line. The resulting transcription factors identified by Garnet were used, together with phosphoproteomic expression changes from the same experimental conditions, as input into Forest. Forest then found a collection of edges from the protein interaction network that connected the two classes of nodes with TGF-β receptors.

The resulting network, depicted in Fig 4, showcases the ability of the forest algorithm to connect known targets (derived from phosphoproteomic and expression data) using the protein-protein interaction network as well as identify hidden 'Steiner' nodes (hexagons) that interact with Garnet-identified transcription factors (triangles) and proteins that exhibit phosphorylation changes (circles). Included among the more robust Steiner nodes (node size correlates with robustness to perturbation) are proteins that have been linked to EMT signaling in cancer, such as PIAS1, a SUMO E3 ligase that is repressed by TGF-β to prevent EMT suppression [56], and COL4A1, which has also been linked to TGF-β stimulation [57]. SUFU and GLI3 have been linked to Hedgehog signaling [58], another pathway in cancer [59], suggesting a putative explanation for the link between TGF-β and Hedgehog signaling. Additional Steiner nodes present in our network, including ABCA1 and ATG12, to the best of our knowledge have not been studied in this context and may point to novel aspects of the TGF-β signaling pathway.

To generate the network illustrated, users can run the test-tgfb-data.py script in the example/a549 directory of the OmicsIntegrator package. This script will run Garnet using the ENCODE-derived DNase I hypersensitive data from A549 cells with the gene expression data from related cells. This script then runs Forest using a scored version of the iRefIndex interactome (version 13) [9] provided with Omics Integrator to identify links between the Garnet transcription factors and the proteins phosphorylated upon TGF-β stimulation.

Table 1. Description of the parameters used in the Forest.py script.

-p PRIZEFILE, --prize=PRIZEFILE	Path to the text file containing the prizes	Path to a tab-delimited plain text file with lines "ProteinName PrizeValue"
-e EDGEFILE, --edge=EDGEFILE	Path to the text file containing the interactome edges	Path to a tab-delimited plain text file with 3 or 4 columns: "ProteinA ProteinB Weight(between 0 and 1) Directionality(U or D, optional)
-c CONFFILE, --conf=CONFFILE	Path to the text file containing the parameters. Default = "./conf.txt"	Path to a tab-delimited plain text file with lines "ParameterName = ParameterValue". Must contain values for w, b, D. Optional parameters mu, garnetBeta, g may also be included
-d DUMMYMODE, --dummyMode=DUMMYMODE	Tells the program which nodes in the interactomes to connect to the dummy root node. Default = "terminals"	Either a file name (containing a list of nodes), "terminals" (connect to all prize nodes), "others" (connect to nodes with no prize), or "all" (connect to all nodes)
--garnet=GARNETOUTPUT	Tells the program that it will also use the Garnet output for network modeling. The prizes will be scaled by the garnetBeta parameter you provide in the conf file, default 0.01	Full path + filename of the Garnet output file
--musquared	Flag to add negative prizes to hub nodes proportional to their degree^2, rather than degree. Use to penalize hub nodes more intensely. Must specify a positive mu in conf file.	
--msgpath=MSGPATH	Path to the msgsteiner executable, including the executable name. Default = "./msgsteiner"	Path where the msgsteiner executable is located
--outpath=OUTPUTPATH	Path to the directory which will hold the output file. Default = this directory	Path
--outlabel=OUTPUTLABEL	A string to put at the beginning of the names of the files output by this program. Default = "result"	String
--cyto30	Use this flag if you want the output files to be compatible with Cytoscape v3.0 (this is the default)	
--cyto28	Use this flag if you want the output files to be compatible with Cytoscape v2.8	
--noisyEdges=NOISENUM	Specifies how many times you would like to add noise to the given edge values and re-run the algorithm. Results of these runs will be merged together and written in files with the word "_noisyEdges_" added to their names. Default = 0	Integer
--shuffledPrizes=SHUFFLENUM	Specifies how many times you would like to shuffle the given prizes around the terminals and re-run the algorithm. Results of these runs will be merged together and written in files with the word "_shuffledPrizes_" added to their names. Default = 0	Integer
--randomTerminals=TERMNUM	Specifies how many times you would like to apply the given prizes to random nodes in the interactome (with a similar degree distribution) and re-run the algorithm. Results of these runs will be merged together and written in files with the word "_randomTerminals_" added to their names. Default = 0	Integer
--knockout=KNOCKOUT	Specifies protein(s) you would like to "knock out" of the interactome to simulate a knock-out experiment	The name(s) of the protein(s), i.e. TP53 or TP53 EGFR
-s SEED, --seed=SEED	A seed for the pseudo-random number generators. If you want to reproduce exact results, supply the same seed. Default = None	Integer

doi:10.1371/journal.pcbi.1004879.t001

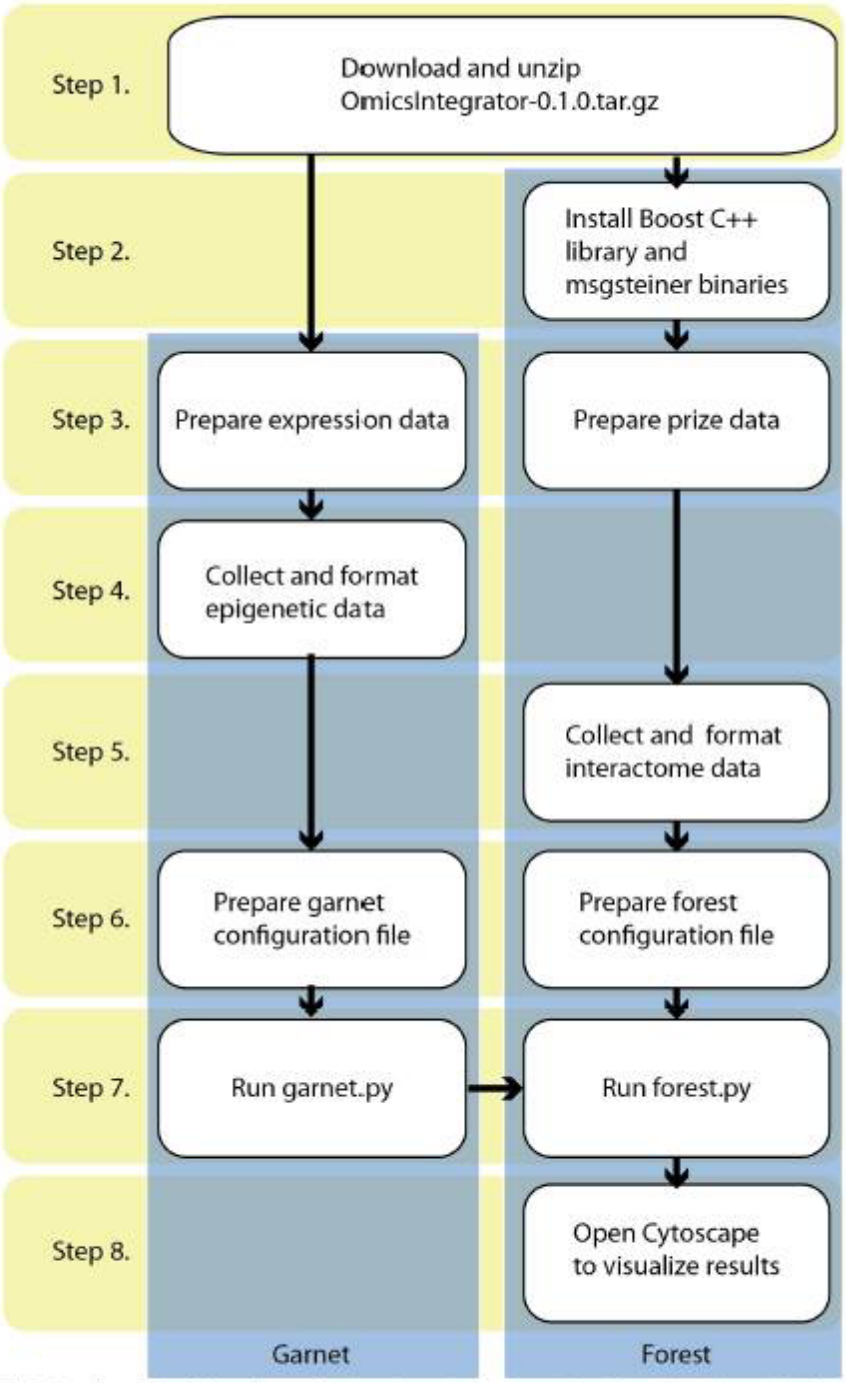


Fig 3. The flowchart of the software. Step 1 requires downloading and unzipping the scripts and data files. Step 2 consists of the installation of the necessary tools to run Omics Integrator. Step 3 describes how to prepare input files. Step 4 and 5 are designed for data collection and formatting for Garnet and Forest

modules, respectively. At Step 6, configuration files are prepared where parameters are defined for Garnet and Forest separately. Garnet and Forest scripts are run at Step 7. If the initial data contains transcriptional data, then Garnet must be run before Forest. Otherwise Forest can be run independently. Detailed instructions of these steps are in the 'Procedure' section of the S1 Text.

doi:10.1371/journal.pcbi.1004879.g003

Network modeling and in silico knock-out experiment in human primary Glioblastoma cells

Forest also has the ability to perform *in silico* knock-outs of specific nodes to model loss or knock-out of these species from a system. Such knockouts can be useful to examine signaling that can occur after a receptor has been inactivated through mutation or pharmacological

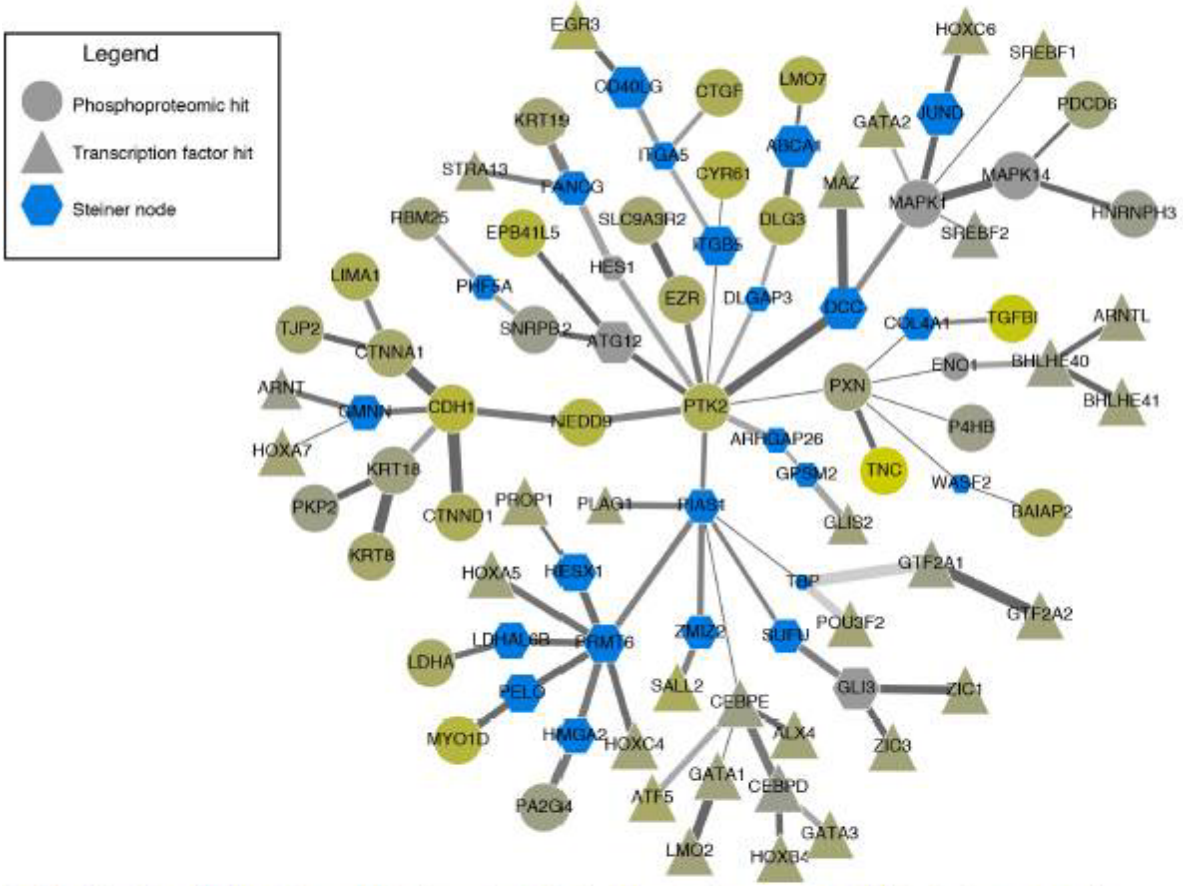


Fig 4. Anticipated results: Network reconstructed from changes in phosphoproteomic measurements (circles) and gene expression measurements (triangles) in lung cancer cell lines stimulated with Tgf-β. Blue hexagons represent 'Steiner nodes' that were not measured as changing in the original experimental measurements but identified through network reconstruction. Nodes that are not blue were measured in the phosphoproteomic data, with color indicating the degree of change in phosphoproteomic measurements: grey indicates no change and yellow indicates a large amount of change. Network robustness was measured by adding noise to the edges using the --noisyEdges flag. The shade of the edge is correlated with the number of times the edge was selected over all perturbations, and the size of a node represents number of times the node was selected. The width of the edge represents the weight assigned to the interaction in the original interactome.

doi:10.1371/journal.pcbi.1004879.g004

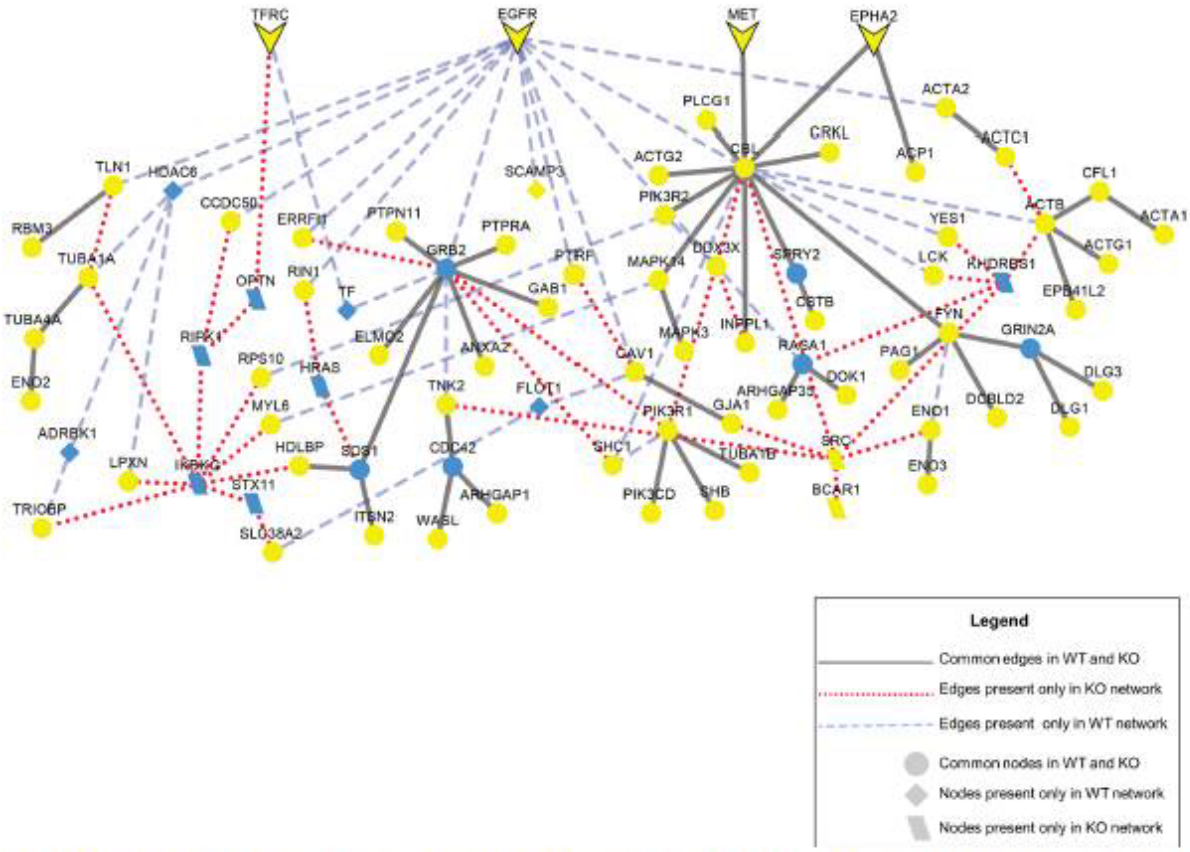


Fig 5. Anticipated results: In silico EGFR knock-out experiment in network modeling. Blue nodes represent 'Steiner nodes' that were not measured as changing in the original experiment but are identified through network reconstruction; yellow nodes represent 'terminal nodes' that are the phosphoproteomic hits. The original network and the network with EGFR knock-out have been merged to clearly show the common and different nodes and edges in the two conditions. Common edges in two conditions are black lines, edges only present in EGFR knock-out condition are red dotted lines and edges only present in the wild-type condition are blue dashed lines. Cell surface receptors are arrow-shaped. The parameters are $\mu = 0.002$, $\omega = 2$, $\beta = 150$, and $D = 10$.

doi:10.1371/journal.pcbi.1004879.g005

treatment. For example, the epidermal growth factor receptor (EGFR) is the target of tyrosine kinase inhibitors including erlotinib and gefitinib. To determine the alternate signaling pathways that could function in the presence of such inhibitors, we use the Forest algorithm to construct a network from phosphoproteomic data measured in U87 cell lines, a model of glioblastoma tumors which has been published previously [60]. We compared this network, called the wild type (WT) network, with a second network built from the same data, but without EGFR included in the interactome (knock-out or KO network). A comparison of the WT and KO networks is depicted in Fig 5. When EGFR is removed from the interactome, the blue dashed edges are removed in the final network, but many key signaling nodes, such as GRB2, CBL and PIK3R1, remain. Also, several cell surface receptors, such as MET, TFRC, EPHA2 are robust to EGFR removal. The network suggests that these receptors could continue signaling to the same downstream targets of EGFR. Indeed, crosstalk between MET and EGFR has been previously identified [61]. These results suggest the presence of alternate pathways that could contribute to the failure of some glioblastoma tumors to respond to treatment.

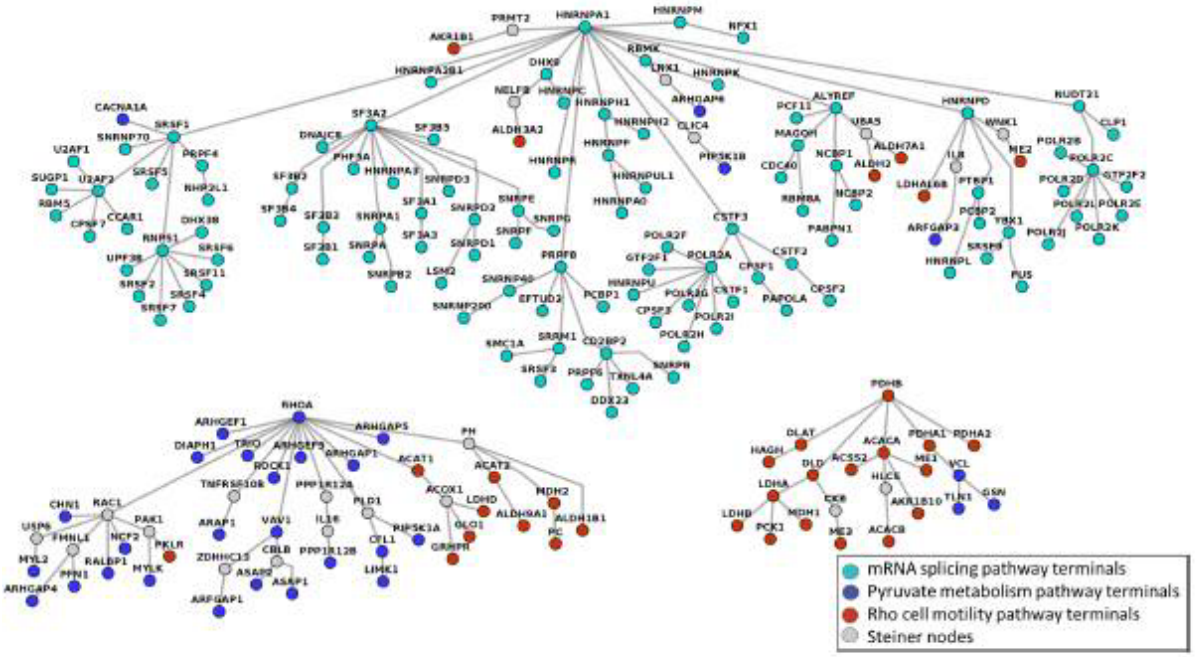


Fig 6. The final PCSF reconstructed from the terminal set formed by the members of mRNA splicing pathway, pyruvate metabolism pathway, and Rho cell motility pathway in ConsensusPathDB. Each node is colored according to the pathway to which it belongs, and Steiner nodes are colored gray. The parameters are $\mu = 0.009$, $\omega = 3$, $\beta = 5$, and $D = 5$.

doi:10.1371/journal.pcbi.1004879.g006

Negative prize weighting reduces bias in interaction networks

Negative prizes can improve the accuracy of Forest's reconstructed networks. We tested the efficacy of adding negative prizes by testing the ability of Forest to reconstruct annotated pathway data from ConsensusPathDB as separate trees [62]. We collected proteins from three pathways in ConsensusPathDB: mRNA splicing, pyruvate metabolism and Rho cell motility. In theory, we would expect these proteins to form three independent trees using the Forest algorithm because they are biologically distinct processes. Without negative prizes, however, Forest assembles all nodes into a single tree (see Fig A in S1 Text). When we penalize high-degree nodes using negative prizes using the μ parameter, the nodes from each pathway form distinct subtrees, illustrated in Fig 6. In addition to recapitulating annotated pathways, nodes in each subtree are enriched for distinct GO processes and shown in Fig B in S1 Text.

Availability and Future Directions

Omics Integrator is an open-source project licensed under the Creative Commons Attribution-NonCommercial 4.0 International Public License. It is available for download at http://fraenkel.mit.edu/omicsintegrator. The install instructions will ensure that all required Python libraries are installed. To solve the prize-collecting Steiner forest problem, Forest uses the message-passing algorithm msgsteiner, which requires a C++ compiler and the Boost library (www.boost.org). The msgsteiner source code can be downloaded from http://areeweb.polito.it/ricerca/cmp/code/bpsteiner. For installation, follow the guidelines in the downloaded files. Development of Omics Integrator is ongoing through our GitHub site (https://github.com/fraenkel-lab/OmicsIntegrator), which provides a framework for collaborations across institutions.

Future directions include adding support for more types of interactions (such as protein-RNA interactions), multiple sets of prizes derived from patient data [32], and additional operating systems. We also plan to improve the parameter selection process and Garnet execution time.

Supporting Information

S1 Text. Supplementary material. Detailed procedure to run Omics Integrator Software and interpret the results.
(DOCX)

Acknowledgments

We thank all the beta testers for running the software, in particular Jonathon Gulliver for his evaluation of the test procedures.

Author Contributions

Conceived and designed the experiments: SJCG NT AG AK EF. Performed the experiments: SJCG NT. Analyzed the data: SJCG NT. Contributed reagents/materials/analysis tools: SJCG NT ARS AK. Wrote the paper: SJCG NT AG AK ARS EF. Designed, developed scripts: NT SJCG ARS AK.

References

1. Yeger-Lotem E, Riva L, Su LJ, Gitler AD, Cashikar AG, et al. (2009) Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. Nat Genet 41: 316–323. doi: 10.1038/ng.337 PMID: 19234470
2. Huang SS, Fraenkel E (2009) Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. Sci Signal 2: ra40. doi: 10.1126/scisignal.2000350 PMID: 19638617
3. Gitter A, Carmi M, Barkai N, Bar-Joseph Z (2013) Linking the signaling cascades and dynamic regulatory networks controlling stress responses. Genome Res 23: 365–376. doi: 10.1101/gr.138628.112 PMID: 23064748
4. Gosline SJ, Spencer SJ, Ursu O, Fraenkel E (2012) SAMNet: a network-based approach to integrate multi-dimensional high throughput datasets. Integr Biol (Camb) 4: 1415–1427.
5. Tuncbag N, Braunstein A, Pagnani A, Huang SS, Chayes J, et al. (2013) Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem. J Comput Biol 20: 124–136. doi: 10.1089/cmb.2012.0092 PMID: 23383998
6. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, 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. PMID: 16199517
7. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4: 44–57. doi: 10.1038/nprot.2008.211 PMID: 19131956
8. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 10: 48. doi: 10.1186/1471-2105-10-48 PMID: 19192299
9. Razick S, Magklaras G, Donaldson IM (2008) iRefIndex: a consolidated protein interaction database with provenance. BMC Bioinformatics 9: 405. doi: 10.1186/1471-2105-9-405 PMID: 18823568
10. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, et al. (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res 41: D808–815. doi: 10.1093/nar/gks1094 PMID: 23203871
11. Chatr-Aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, et al. (2013) The BioGRID interaction database: 2013 update. Nucleic Acids Res 41: D816–823. doi: 10.1093/nar/gks1158 PMID: 23203989

12. Isserlin R, El-Badrawi RA, Bader GD (2011) The Biomolecular Interaction Network Database in PSI-MI 2.5. Database (Oxford) 2011: baq037.

13. Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, et al. (2013) HMDB 3.0—The Human Metabolome Database in 2013. Nucleic Acids Res 41: D801–807. doi: 10.1093/nar/gks1065 PMID: 23161693

14. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, et al. (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 36: D901–906. PMID: 18048412

15. Kuhn M, Szklarczyk D, Pletscher-Frankild S, Blicher TH, von Mering C, et al. (2014) STITCH 4: integration of protein-chemical interactions with user data. Nucleic Acids Res 42: D401–407. doi: 10.1093/nar/gkt1207 PMID: 24293645

16. Furey TS (2012) ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat Rev Genet 13: 840–852. doi: 10.1038/nrg3306 PMID: 23090257

17. Alcaraz N, Pauling J, Batra R, Barbosa E, Junge A, et al. (2014) KeyPathwayMiner 4.0: condition-specific pathway analysis by combining multiple omics studies and networks with Cytoscape. BMC Syst Biol 8: 99. doi: 10.1186/s12918-014-0099-x PMID: 25134827

18. Conlon EM, Liu XS, Lieb JD, Liu JS (2003) Integrating regulatory motif discovery and genome-wide expression analysis. Proc Natl Acad Sci U S A 100: 3339–3344. PMID: 12626739

19. Gitter A, Bar-Joseph Z (2013) Identifying proteins controlling key disease signaling pathways. Bioinformatics 29: i227–236. doi: 10.1093/bioinformatics/btt241 PMID: 23812988

20. Patil A, Kumagai Y, Liang KC, Suzuki Y, Nakai K (2013) Linking transcriptional changes over time in stimulated dendritic cells to identify gene networks activated during the innate immune response. PLoS Comput Biol 9: e1003323. doi: 10.1371/journal.pcbi.1003323 PMID: 24244133

21. Paull EO, Carlin DE, Niepel M, Sorger PK, Haussler D, et al. (2013) Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). Bioinformatics 29: 2757–2764. doi: 10.1093/bioinformatics/btt471 PMID: 23986566

22. Pique-Regi R, Degner JF, Pai AA, Gaffney DJ, Gilad Y, et al. (2011) Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res 21: 447–455. doi: 10.1101/gr.112623.110 PMID: 21106904

23. Roven C, Bussemaker HJ (2003) REDUCE: An online tool for inferring cis-regulatory elements and transcriptional module activities from microarray data. Nucleic Acids Res 31: 3487–3490. PMID: 12824350

24. Sherwood RI, Hashimoto T, O'Donnell CW, Lewis S, Barkal AA, et al. (2014) Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. Nat Biotechnol 32: 171–178. doi: 10.1038/nbt.2798 PMID: 24441470

25. Vandin F, Upfal E, Raphael BJ (2011) Algorithms for detecting significantly mutated pathways in cancer. J Comput Biol 18: 507–522. doi: 10.1089/cmb.2010.0265 PMID: 21385051

26. Gitter A, Klein-Seetharaman J, Gupta A, Bar-Joseph Z (2011) Discovering pathways by orienting edges in protein interaction networks. Nucleic Acids Res 39: e22. doi: 10.1093/nar/gkq1207 PMID: 21109539

27. Ourfali O, Shlomi T, Ideker T, Ruppin E, Sharan R (2007) SPINE: a framework for signaling-regulatory pathway inference from cause-effect experiments. Bioinformatics 23: i359–366. PMID: 17646318

28. Lan A, Smoly IY, Rapaport G, Lindquist S, Fraenkel E, et al. (2011) ResponseNet: revealing signaling and regulatory networks linking genetic and transcriptomic screening data. Nucleic Acids Res 39: W424–429. doi: 10.1093/nar/gkr359 PMID: 21576238

29. Chasman D, Ho YH, Berry DB, Nemec CM, MacGilvray ME, et al. (2014) Pathway connectivity and signaling coordination in the yeast stress-activated signaling network. Mol Syst Biol 10: 759. doi: 10.15252/msb.20145120 PMID: 25411400

30. Ulitsky I, Shamir R (2009) Identifying functional modules using expression profiles and confidence-scored protein interactions. Bioinformatics 25: 1158–1164. doi: 10.1093/bioinformatics/btp118 PMID: 19297352

31. Cheng C, Yan KK, Yip KY, Rozowsky J, Alexander R, et al. (2011) A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. Genome Biol 12: R15. doi: 10.1186/gb-2011-12-2-r15 PMID: 21324173

32. Gitter A, Braunstein A, Pagnani A, Baldassi C, Borgs C, et al. (2014) Sharing information to reconstruct patient-specific pathways in heterogeneous diseases. Pac Symp Biocomput: 39–50. doi: http://dx.doi.org/10.1142/9789814583220_0005 PMID: 24297532

33. Gosline SJ, Oh C, Fraenkel E (2014) SAMNetWeb: identifying condition-specific networks linking signaling and transcription. Bioinformatics. PMID: 25414365

34. Consortium EP (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 306: 636–640. PMID: 15499007

35. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, et al. (2015) Integrative analysis of 111 reference human epigenomes. Nature 518: 317–330. doi: 10.1038/nature14248 PMID: 25693563

36. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, et al. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 34: D108–110. PMID: 16381825

37. Ng CW, Yildirim F, Yap YS, Dalin S, Matthews BJ, et al. (2013) Extensive changes in DNA methylation are associated with expression of mutant huntingtin. Proc Natl Acad Sci U S A 110: 2354–2359. doi: 10.1073/pnas.1221292110 PMID: 23341638

38. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498–2504. PMID: 14597658

39. Hakes L, Pinney JW, Robertson DL, Lovell SC (2008) Protein-protein interaction networks and biology—what's the connection? Nat Biotechnol 26: 69–72. doi: 10.1038/nbt0108-69 PMID: 18183023

40. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, et al. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417: 399–403. PMID: 12000970

41. Aloy P, Russell RB (2002) Potential artefacts in protein-interaction networks. FEBS Lett 530: 253–254. PMID: 12387902

42. Thomson S, Petti F, Sujka-Kwok I, Mercado P, Bean J, et al. (2011) A systems view of epithelial-mesenchymal transition signaling states. Clin Exp Metastasis 28: 137–155. doi: 10.1007/s10585-010-9367-3 PMID: 21194007

43. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, et al. (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483: 603–607. doi: 10.1038/nature11003 PMID: 22460905

44. Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, et al. (2011) Global quantification of mammalian gene expression control. Nature 473: 337–342. doi: 10.1038/nature10098 PMID: 21593866

45. Vogel C, Marcotte EM (2012) Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet 13: 227–232. doi: 10.1038/nrg3185 PMID: 22411467

46. Li JJ, Biggin MD (2015) Gene expression. Statistics requantitates the central dogma. Science 347: 1066–1067. doi: 10.1126/science.aaa8332 PMID: 25745146

47. Huang SS, Clarke DC, Gosline SJ, Labadorf A, Chouinard CR, et al. (2013) Linking proteomic and transcriptional data through the interactome and epigenome reveals a map of oncogene-induced signaling. PLoS Comput Biol 9: e1002887. doi: 10.1371/journal.pcbi.1002887 PMID: 23408876

48. Gosline SJ, Gurtan AM, JnBaptiste CK, Bosson A, Milani P, et al. (2016) Elucidating MicroRNA Regulatory Networks Using Transcriptional, Post-transcriptional, and Histone Modification Measurements. Cell Rep 14: 310–319. doi: 10.1016/j.celrep.2015.12.031 PMID: 26748710

49. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, et al. (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. Nature 489: 83–90. doi: 10.1038/nature11212 PMID: 22955618

50. Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, et al. (2011) Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res 21: 1757–1767. doi: 10.1101/gr.121541.111 PMID: 21750106

51. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, et al. (2012) MINT, the molecular interaction database: 2012 update. Nucleic Acids Res 40: D857–861. doi: 10.1093/nar/gkr930 PMID: 22096227

52. Suthram S, Shlomi T, Ruppin E, Sharan R, Ideker T (2006) A direct comparison of protein interaction confidence assignment schemes. BMC Bioinformatics 7: 360. PMID: 16872496

53. Bailly-Bechet M, Borgs C, Braunstein A, Chayes J, Dagkessamanskaia A, et al. (2011) Finding undetected protein associations in cell signaling by belief propagation. Proc Natl Acad Sci U S A 108: 882–887. doi: 10.1073/pnas.1004751108 PMID: 21187432

54. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, et al. (2011) Molecular signatures database (MSigDB) 3.0. Bioinformatics 27: 1739–1740. doi: 10.1093/bioinformatics/btr260 PMID: 21546393

55. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29. PMID: 10802651

56. Netherton SJ, Bonni S (2010) Suppression of TGFbeta-induced epithelial-mesenchymal transition like phenotype by a PIAS1 regulated sumoylation pathway in NMuMG epithelial cells. PLoS One 5: e13971. doi: 10.1371/journal.pone.0013971 PMID: 21103059

57. Agajanian M, Runa F, Kelber JA (2015) Identification of a PEAK1/ZEB1 signaling axis during TGFbeta/fibronectin-induced EMT in breast cancer. Biochem Biophys Res Commun 465: 606–612. doi: 10.1016/j.bbrc.2015.08.071 PMID: 26297948

58. Humke EW, Dorn KV, Milenkovic L, Scott MP, Rohatgi R (2010) The output of Hedgehog signaling is controlled by the dynamic association between Suppressor of Fused and the Gli proteins. Genes Dev 24: 670–682. doi: 10.1101/gad.1902910 PMID: 20360384

59. Chatel G, Ganeff C, Boussif N, Delacroix L, Briquet A, et al. (2007) Hedgehog signaling pathway is inactive in colorectal cancer cell lines. Int J Cancer 121: 2622–2627. PMID: 17683069

60. Huang PH, Mukasa A, Bonavia R, Flynn RA, Brewer ZE, et al. (2007) Quantitative analysis of EGFRvIII cellular signaling networks reveals a combinatorial therapeutic strategy for glioblastoma. Proc Natl Acad Sci U S A 104: 12867–12872. PMID: 17646646

61. Velpula KK, Dasari VR, Asuthkar S, Gorantla B, Tsung AJ (2012) EGFR and c-Met Cross Talk in Glioblastoma and Its Regulation by Human Cord Blood Stem Cells. Transl Oncol 5: 379–392. PMID: 23066446

62. Kamburov A, Pentchev K, Galicka H, Wierling C, Lehrach H, et al. (2011) ConsensusPathDB: toward a more complete picture of cell biology. Nucleic Acids Res 39: D712–717. doi: 10.1093/nar/gkq1156 PMID: 21071422

REFERENCES

1. *Overweight and obesity statistics* N.I.o.D.a.D.a.K.D. (NIDDK), Editor 2012, National Institutes of Health.
2. Ogden, C.L., et al., *Prevalence of childhood and adult obesity in the United States, 2011-2012.* JAMA, 2014. **311**(8): p. 806-14.
3. Ogden, C.L., Carroll, M.D., Fryar, C.D., Flegal, K.M., *Prevalence of obesity among adults and youth: United States, 2011-2014.* 2015. **NCHS Data Brief**(219).
4. *Obesity and overweight*, W.H.O. (WHO), Editor 2016.
5. Flegal, K.M., et al., *Association of all-cause mortality with overweight and obesity using standard body mass index categories: a systematic review and meta-analysis.* JAMA, 2013. **309**(1): p. 71-82.
6. Eckel, R.H., S.M. Grundy, and P.Z. Zimmet, *The metabolic syndrome.* Lancet, 2005. **365**(9468): p. 1415-28.
7. Lusis, A.J., A.D. Attie, and K. Reue, *Metabolic syndrome: from epidemiology to systems biology.* Nature reviews. Genetics, 2008. **9**(11): p. 819-30.
8. Kahn, S.E., R.L. Hull, and K.M. Utzschneider, *Mechanisms linking obesity to insulin resistance and type 2 diabetes.* Nature, 2006. **444**(7121): p. 840-6.
9. Boyle, J.P., et al., *Projection of the year 2050 burden of diabetes in the US adult population: dynamic modeling of incidence, mortality, and prediabetes prevalence.* Population health metrics, 2010. **8**: p. 29.
10. Scherer, P.E., *Adipose tissue: from lipid storage compartment to endocrine organ.* Diabetes, 2006. **55**(6): p. 1537-45.
11. Fain, J.N., et al., *Comparison of the release of adipokines by adipose tissue, adipose tissue matrix, and adipocytes from visceral and subcutaneous abdominal adipose tissues of obese humans.* Endocrinology, 2004. **145**(5): p. 2273-82.
12. Shoelson, S.E., L. Herrero, and A. Naaz, *Obesity, inflammation, and insulin resistance.* Gastroenterology, 2007. **132**(6): p. 2169-80.
13. Kovacs, P. and M. Stumvoll, *Fatty acids and insulin resistance in muscle and liver.* Best practice & research. Clinical endocrinology & metabolism, 2005. **19**(4): p. 625-35.
14. Krssak, M., et al., *Intramyocellular lipid concentrations are correlated with insulin sensitivity in humans: a 1H NMR spectroscopy study.* Diabetologia, 1999. **42**(1): p. 113-6.
15. DeFronzo, R.A., et al., *Effects of insulin on peripheral and splanchnic glucose metabolism in noninsulin-dependent (type II) diabetes mellitus.* The Journal of clinical investigation, 1985. **76**(1): p. 149-55.
16. Kahn, S.E., *Clinical review 135: The importance of beta-cell failure in the development and progression of type 2 diabetes.* The Journal of clinical endocrinology and metabolism, 2001. **86**(9): p. 4047-58.
17. Kahn, S.E., M.E. Cooper, and S. Del Prato, *Pathophysiology and treatment of type 2 diabetes: perspectives on the past, present, and future.* Lancet, 2014. **383**(9922): p. 1068-83.
18. Knowles, N.G., et al., *Insulin and amylin release are both diminished in first-degree relatives of subjects with type 2 diabetes.* Diabetes care, 2002. **25**(2): p. 292-7.
19. McCarthy, M.I., *Genomics, type 2 diabetes, and obesity.* The New England journal of medicine, 2010. **363**(24): p. 2339-50.
20. Nathan, D.M., et al., *Medical management of hyperglycemia in type 2 diabetes: a consensus algorithm for the initiation and adjustment of therapy: a consensus statement of the American Diabetes Association and the European Association for the Study of Diabetes.* Diabetes care, 2009. **32**(1): p. 193-203.
21. Palmer, S.C., et al., *Comparison of Clinical Outcomes and Adverse Events Associated With Glucose-Lowering Drugs in Patients With Type 2 Diabetes: A Meta-analysis.* JAMA, 2016. **316**(3): p. 313-24.
22. *(7) Approaches to glycemic treatment.* Diabetes care, 2015. **38 Suppl**: p. S41-8.
23. Maruthur, N.M., et al., *Diabetes Medications as Monotherapy or Metformin-Based Combination Therapy for Type 2 Diabetes A Systematic Review and Meta-analysis.* Annals of Internal Medicine, 2016. **164**(11): p. 740-+.
24. Viollet, B. and M. Foretz, *Revisiting the mechanisms of metformin action in the liver.* Annales d'endocrinologie, 2013. **74**(2): p. 123-9.
25. Pfeiffer, A.F. and H.H. Klein, *The treatment of type 2 diabetes.* Deutsches Arzteblatt international, 2014. **111**(5): p. 69-81; quiz 82.
26. Bischoff, H., *The mechanism of alpha-glucosidase inhibition in the management of diabetes.* Clinical and investigative medicine. Medecine clinique et experimentale, 1995. **18**(4): p. 303-11.
27. Yki-Jarvinen, H., *Thiazolidinediones.* The New England journal of medicine, 2004. **351**(11): p. 1106-18.

28. DeFronzo, R.A., D. Simonson, and E. Ferrannini, *Hepatic and peripheral insulin resistance: a common feature of type 2 (non-insulin-dependent) and type 1 (insulin-dependent) diabetes mellitus.* Diabetologia, 1982. **23**(4): p. 313-9.
29. Samuel, V.T. and G.I. Shulman, *Mechanisms for insulin resistance: common threads and missing links.* Cell, 2012. **148**(5): p. 852-71.
30. Pilkis, S.J. and D.K. Granner, *Molecular physiology of the regulation of hepatic gluconeogenesis and glycolysis.* Annual review of physiology, 1992. **54**: p. 885-909.
31. He, L., et al., *Metformin and insulin suppress hepatic gluconeogenesis through phosphorylation of CREB binding protein.* Cell, 2009. **137**(4): p. 635-46.
32. Taniguchi, C.M., B. Emanuelli, and C.R. Kahn, *Critical nodes in signalling pathways: insights into insulin action.* Nature reviews. Molecular cell biology, 2006. **7**(2): p. 85-96.
33. Perry, R.J., et al., *The role of hepatic lipids in hepatic insulin resistance and type 2 diabetes.* Nature, 2014. **510**(7503): p. 84-91.
34. Park, S., et al., *Effect of high fat diet on insulin resistance: dietary fat versus visceral fat mass.* Journal of Korean medical science, 2001. **16**(4): p. 386-90.
35. Samuel, V.T., et al., *Mechanism of hepatic insulin resistance in non-alcoholic fatty liver disease.* The Journal of biological chemistry, 2004. **279**(31): p. 32345-53.
36. Hirosumi, J., et al., *A central role for JNK in obesity and insulin resistance.* Nature, 2002. **420**(6913): p. 333-6.
37. Aguirre, V., et al., *The c-Jun NH(2)-terminal kinase promotes insulin resistance during association with insulin receptor substrate-1 and phosphorylation of Ser(307).* The Journal of biological chemistry, 2000. **275**(12): p. 9047-54.
38. Samuel, V.T., et al., *Inhibition of protein kinase Cepsilon prevents hepatic insulin resistance in nonalcoholic fatty liver disease.* The Journal of clinical investigation, 2007. **117**(3): p. 739-45.
39. Shimomura, I., et al., *Decreased IRS-2 and increased SREBP-1c lead to mixed insulin resistance and sensitivity in livers of lipodystrophic and ob/ob mice.* Molecular cell, 2000. **6**(1): p. 77-86.
40. Brown, M.S. and J.L. Goldstein, *Selective versus total insulin resistance: a pathogenic paradox.* Cell metabolism, 2008. **7**(2): p. 95-6.
41. Michael, M.D., et al., *Loss of insulin signaling in hepatocytes leads to severe insulin resistance and progressive hepatic dysfunction.* Molecular cell, 2000. **6**(1): p. 87-97.
42. Fisher, S.J. and C.R. Kahn, *Insulin signaling is required for insulin's direct and indirect action on hepatic glucose production.* The Journal of clinical investigation, 2003. **111**(4): p. 463-8.
43. Kitamura, T., C.R. Kahn, and D. Accili, *Insulin receptor knockout mice.* Annual review of physiology, 2003. **65**: p. 313-32.
44. Buettner, C., et al., *Severe impairment in liver insulin signaling fails to alter hepatic insulin action in conscious mice.* The Journal of clinical investigation, 2005. **115**(5): p. 1306-13.
45. Lu, M., et al., *Insulin regulates liver metabolism in vivo in the absence of hepatic Akt and Foxo1.* Nature medicine, 2012. **18**(3): p. 388-95.
46. Titchenell, P.M., et al., *Hepatic insulin signalling is dispensable for suppression of glucose output by insulin in vivo.* Nature communications, 2015. **6**: p. 7078.
47. Perry, R.J., et al., *Hepatic acetyl CoA links adipose tissue inflammation to hepatic insulin resistance and type 2 diabetes.* Cell, 2015. **160**(4): p. 745-58.
48. Titchenell, P.M., et al., *Direct Hepatocyte Insulin Signaling Is Required for Lipogenesis but Is Dispensable for the Suppression of Glucose Production.* Cell metabolism, 2016. **23**(6): p. 1154-66.
49. Kanuri, G. and I. Bergheim, *In vitro and in vivo models of non-alcoholic fatty liver disease (NAFLD).* International journal of molecular sciences, 2013. **14**(6): p. 11963-80.
50. Wobser, H., et al., *Lipid accumulation in hepatocytes induces fibrogenic activation of hepatic stellate cells.* Cell research, 2009. **19**(8): p. 996-1005.
51. Chabowski, A., et al., *Fatty acid transporters involved in the palmitate and oleate induced insulin resistance in primary rat hepatocytes.* Acta Physiologica, 2013. **207**(2): p. 346-357.
52. Lo, K.A., et al., *Analysis of in vitro insulin-resistance models and their physiological relevance to in vivo diet-induced adipose insulin resistance.* Cell reports, 2013. **5**(1): p. 259-70.
53. Ishii, M., et al., *Palmitate induces insulin resistance in human HepG2 hepatocytes by enhancing ubiquitination and proteasomal degradation of key insulin signaling molecules.* Archives of biochemistry and biophysics, 2015. **566**: p. 26-35.

54. Hart, S.N., et al., *A comparison of whole genome gene expression profiles of HepaRG cells and HepG2 cells to primary human hepatocytes and human liver tissues.* Drug metabolism and disposition: the biological fate of chemicals, 2010. **38**(6): p. 988-94.
55. Kanasaki, K. and D. Koya, *Biology of obesity: lessons from animal models of obesity.* Journal of biomedicine & biotechnology, 2011. **2011**: p. 197636.
56. Radonjic, M., et al., *Genome-wide mRNA expression analysis of hepatic adaptation to high-fat diets reveals switch from an inflammatory to steatotic transcriptional program.* PLoS One, 2009. **4**(8): p. e6646.
57. Clee, S.M. and A.D. Attie, *The genetic landscape of type 2 diabetes in mice.* Endocrine reviews, 2007. **28**(1): p. 48-83.
58. Wang, B., P.C. Chandrasekera, and J.J. Pippin, *Leptin- and leptin receptor-deficient rodent models: relevance for human type 2 diabetes.* Current diabetes reviews, 2014. **10**(2): p. 131-45.
59. Chalasani, N., et al., *Does leptin play a role in the pathogenesis of human nonalcoholic steatohepatitis?* The American journal of gastroenterology, 2003. **98**(12): p. 2771-6.
60. Surwit, R.S., et al., *Diet-induced type II diabetes in C57BL/6J mice.* Diabetes, 1988. **37**(9): p. 1163-7.
61. Winzell, M.S. and B. Ahren, *The high-fat diet-fed mouse: a model for studying mechanisms and treatment of impaired glucose tolerance and type 2 diabetes.* Diabetes, 2004. **53 Suppl 3**: p. S215-9.
62. Collins, S., et al., *Genetic vulnerability to diet-induced obesity in the C57BL/6J mouse: physiological and molecular characteristics.* Physiology & behavior, 2004. **81**(2): p. 243-8.
63. Zhao, Y., R.E. Barrere-Cain, and X. Yang, *Nutritional systems biology of type 2 diabetes.* Genes & nutrition, 2015. **10**(5): p. 481.
64. Sanghera, D.K. and P.R. Blackett, *Type 2 Diabetes Genetics: Beyond GWAS.* Journal of diabetes & metabolism, 2012. **3**(198).
65. Kitano, H., *Systems biology: a brief overview.* Science, 2002. **295**(5560): p. 1662-4.
66. Bhalla, U.S. and R. Iyengar, *Emergent properties of networks of biological signaling pathways.* Science, 1999. **283**(5400): p. 381-7.
67. Sauer, U., M. Heinemann, and N. Zamboni, *Genetics. Getting closer to the whole picture.* Science, 2007. **316**(5824): p. 550-1.
68. Bertalanffy, L.v., *General system theory; foundations, development, applications*. Rev. ed1969, New York,: G. Braziller. xxiv, 295 p.
69. Hodgkin, A.L. and A.F. Huxley, *A quantitative description of membrane current and its application to conduction and excitation in nerve.* The Journal of physiology, 1952. **117**(4): p. 500-44.
70. Noble, D., *Cardiac action and pacemaker potentials based on the Hodgkin-Huxley equations.* Nature, 1960. **188**: p. 495-7.
71. Luo, C.H. and Y. Rudy, *A model of the ventricular cardiac action potential. Depolarization, repolarization, and their interaction.* Circulation Research, 1991. **68**(6): p. 1501-1526.
72. Shannon, T.R., et al., *A mathematical treatment of integrated Ca dynamics within the ventricular myocyte.* Biophysical journal, 2004. **87**(5): p. 3351-71.
73. Omenn, G.S., *Grand challenges and great opportunities in science, technology, and public policy.* Science, 2006. **314**(5806): p. 1696-1704.
74. Hunter, P., *Back down to Earth. Even if it has not yet lived up to its promises, systems biology has now matured and is about to deliver its first results.* EMBO reports, 2012. **13**(5): p. 408-11.
75. Yildirim, N. and M.C. Mackey, *Feedback regulation in the lactose operon: a mathematical modeling study and comparison with experimental data.* Biophysical journal, 2003. **84**(5): p. 2841-51.
76. Thiele, I., et al., *A community-driven global reconstruction of human metabolism.* Nature biotechnology, 2013. **31**(5): p. 419-25.
77. Karr, J.R., et al., *A whole-cell computational model predicts phenotype from genotype.* Cell, 2012. **150**(2): p. 389-401.
78. Meng, Q., et al., *Systems Biology Approaches and Applications in Obesity, Diabetes, and Cardiovascular Diseases.* Current cardiovascular risk reports, 2013. **7**(1): p. 73-83.
79. Horgan, R.P. and L.C. Kenny, *'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics.* The Obstetrician & Gynaecologist, 2011. **13**(3): p. 189-195.
80. Metzker, M.L., *Sequencing technologies - the next generation.* Nature reviews. Genetics, 2010. **11**(1): p. 31-46.
81. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nature methods, 2008. **5**(7): p. 621-8.

82. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nature reviews. Genetics, 2009. **10**(1): p. 57-63.
83. Zhang, W., et al., *Comparison of RNA-seq and microarray-based models for clinical endpoint prediction.* Genome biology, 2015. **16**: p. 133.
84. Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nature biotechnology, 2010. **28**(5): p. 511-5.
85. Katz, Y., et al., *Analysis and design of RNA sequencing experiments for identifying isoform regulation.* Nature methods, 2010. **7**(12): p. 1009-15.
86. Shen, S., et al., *rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data.* Proceedings of the National Academy of Sciences, 2014. **111**(51): p. E5593-E5601.
87. Ekblom, R. and J.B. Wolf, *A field guide to whole-genome sequencing, assembly and annotation.* Evolutionary applications, 2014. **7**(9): p. 1026-42.
88. Meyer, C.A. and X.S. Liu, *Identifying and mitigating bias in next-generation sequencing methods for chromatin biology.* Nature reviews. Genetics, 2014. **15**(11): p. 709-721.
89. Kundaje, A., et al., *Integrative analysis of 111 reference human epigenomes.* Nature, 2015. **518**(7539): p. 317-30.
90. Santos-Rosa, H., et al., *Active genes are tri-methylated at K4 of histone H3.* Nature, 2002. **419**(6905): p. 407-11.
91. Creyghton, M.P., et al., *Histone H3K27ac separates active from poised enhancers and predicts developmental state.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(50): p. 21931-6.
92. Young, M.D., et al., *ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity.* Nucleic acids research, 2011. **39**(17): p. 7415-27.
93. Strahl, B.D. and C.D. Allis, *The language of covalent histone modifications.* Nature, 2000. **403**(6765): p. 41-5.
94. Park, P.J., *ChIP-seq: advantages and challenges of a maturing technology.* Nature reviews. Genetics, 2009. **10**(10): p. 669-80.
95. Song, L. and G.E. Crawford, *DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells.* Cold Spring Harbor protocols, 2010. **2010**(2): p. pdb prot5384.
96. Simon, J.M., et al., *Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA.* Nat. Protocols, 2012. **7**(2): p. 256-267.
97. Buenrostro, J.D., et al., *Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position.* Nat Meth, 2013. **10**(12): p. 1213-1218.
98. Ng, C.W., et al., *Extensive changes in DNA methylation are associated with expression of mutant huntingtin.* Proceedings of the National Academy of Sciences of the United States of America, 2013. **110**(6): p. 2354-9.
99. Lister, R., et al., *Global epigenomic reconfiguration during mammalian brain development.* Science, 2013. **341**(6146): p. 1237905.
100. Larance, M. and A.I. Lamond, *Multidimensional proteomics for cell biology.* Nature reviews. Molecular cell biology, 2015. **16**(5): p. 269-80.
101. Sandhu, C., et al., *Evaluation of data-dependent versus targeted shotgun proteomic approaches for monitoring transcription factor expression in breast cancer.* Journal of proteome research, 2008. **7**(4): p. 1529-41.
102. Boja, E.S. and H. Rodriguez, *Mass spectrometry-based targeted quantitative proteomics: achieving sensitive and reproducible detection of proteins.* Proteomics, 2012. **12**(8): p. 1093-110.
103. Ross, P.L., et al., *Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents.* Molecular & cellular proteomics : MCP, 2004. **3**(12): p. 1154-69.
104. Cox, J., et al., *Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ.* Molecular & Cellular Proteomics, 2014. **13**(9): p. 2513-2526.
105. Ong, S.-E., et al., *Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics.* Molecular & Cellular Proteomics, 2002. **1**(5): p. 376-386.
106. Schwanhausser, B., et al., *Global quantification of mammalian gene expression control.* Nature, 2011. **473**(7347): p. 337-42.
107. Kim, M.-S., et al., *A draft map of the human proteome.* Nature, 2014. **509**(7502): p. 575-581.

108. Trinkle-Mulcahy, L., et al., *Identifying specific protein interaction partners using quantitative mass spectrometry and bead proteomes.* The Journal of cell biology, 2008. **183**(2): p. 223-39.
109. Roux, K.J., et al., *A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells.* The Journal of cell biology, 2012. **196**(6): p. 801-10.
110. Razick, S., G. Magklaras, and I.M. Donaldson, *iRefIndex: a consolidated protein interaction database with provenance.* BMC bioinformatics, 2008. **9**: p. 405.
111. Nicholson, J.K. and J.C. Lindon, *Systems biology: Metabonomics.* Nature, 2008. **455**(7216): p. 1054-6.
112. McDunn, J.E., et al., *Metabolomic signatures of aggressive prostate cancer.* The Prostate, 2013. **73**(14): p. 1547-60.
113. Chinnaiyan, P., et al., *The metabolomic signature of malignant glioma reflects accelerated anabolic metabolism.* Cancer research, 2012. **72**(22): p. 5878-88.
114. Jansson, J., et al., *Metabolomics reveals metabolic biomarkers of Crohn's disease.* PloS one, 2009. **4**(7): p. e6386.
115. Wang-Sattler, R., et al., *Novel biomarkers for pre-diabetes identified by metabolomics.* Molecular systems biology, 2012. **8**: p. 615.
116. Suhre, K., et al., *Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting.* PloS one, 2010. **5**(11): p. e13953.
117. Wishart, D.S., et al., *HMDB 3.0--The Human Metabolome Database in 2013.* Nucleic acids research, 2013. **41**(Database issue): p. D801-7.
118. Dunn, W.B., et al., *Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy.* Chemical Society reviews, 2011. **40**(1): p. 387-426.
119. Kuczynski, J., et al., *Experimental and analytical tools for studying the human microbiome.* Nature reviews. Genetics, 2012. **13**(1): p. 47-58.
120. Wang, J. and H. Jia, *Metagenome-wide association studies: fine-mining the microbiome.* Nat Rev Micro, 2016. **14**(8): p. 508-522.
121. Sayin, S.I., et al., *Gut microbiota regulates bile acid metabolism by reducing the levels of tauro-beta-muricholic acid, a naturally occurring FXR antagonist.* Cell metabolism, 2013. **17**(2): p. 225-35.
122. Karolchik, D., et al., *The UCSC Table Browser data retrieval tool.* Nucleic acids research, 2004. **32**(Database issue): p. D493-6.
123. Pique-Regi, R., et al., *Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data.* Genome research, 2011. **21**(3): p. 447-55.
124. Cuellar-Partida, G., et al., *Epigenetic priors for identifying active transcription factor binding sites.* Bioinformatics, 2012. **28**(1): p. 56-62.
125. Macisaac, K.D., et al., *A hypothesis-based approach for identifying the binding specificity of regulatory proteins from chromatin immunoprecipitation data.* Bioinformatics, 2006. **22**(4): p. 423-9.
126. Bailey, T.L., et al., *MEME: discovering and analyzing DNA and protein sequence motifs.* Nucleic acids research, 2006. **34**(Web Server issue): p. W369-73.
127. Heinz, S., et al., *Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities.* Molecular cell, 2010. **38**(4): p. 576-89.
128. Bryne, J.C., et al., *JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update.* Nucleic Acids Res, 2008. **36**(Database issue): p. D102-6.
129. Wingender, E., et al., *TRANSFAC: a database on transcription factors and their DNA binding sites.* Nucleic acids research, 1996. **24**(1): p. 238-41.
130. Bersanelli, M., et al., *Methods for the integration of multi-omics data: mathematical aspects.* BMC bioinformatics, 2016. **17 Suppl 2**: p. 15.
131. Berger, B., J. Peng, and M. Singh, *Computational solutions for omics data.* Nature reviews. Genetics, 2013. **14**(5): p. 333-46.
132. Kuo, T.C., T.F. Tian, and Y.J. Tseng, *3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data.* BMC systems biology, 2013. **7**: p. 64.
133. Oberbach, A., et al., *Combined proteomic and metabolomic profiling of serum reveals association of the complement system with obesity and identifies novel markers of body fat mass changes.* Journal of proteome research, 2011. **10**(10): p. 4769-88.
134. Geladi, P. and B.R. Kowalski, *Partial Least-Squares Regression - a Tutorial.* Analytica Chimica Acta, 1986. **185**: p. 1-17.

135. Miraldi, E.R., et al., *Molecular network analysis of phosphotyrosine and lipid metabolism in hepatic PTP1b deletion mice.* Integrative biology : quantitative biosciences from nano to macro, 2013. **5**(7): p. 940-63.
136. Le Cao, K.A., I. Gonzalez, and S. Dejean, *integrOmics: an R package to unravel relationships between two omics datasets.* Bioinformatics, 2009. **25**(21): p. 2855-6.
137. Li, W., et al., *Identifying multi-layer gene regulatory modules from multi-dimensional genomic data.* Bioinformatics, 2012. **28**(19): p. 2458-66.
138. Srivastava, V., et al., *OnPLS integration of transcriptomic, proteomic and metabolomic data shows multi-level oxidative stress responses in the cambium of transgenic hipI- superoxide dismutase Populus plants.* BMC genomics, 2013. **14**: p. 893.
139. Blanchet, L., et al., *Fusion of metabolomics and proteomics data for biomarkers discovery: case study on the experimental autoimmune encephalomyelitis.* BMC bioinformatics, 2011. **12**: p. 254.
140. Ouyang, Z., Q. Zhou, and W.H. Wong, *ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells.* Proceedings of the National Academy of Sciences of the United States of America, 2009. **106**(51): p. 21521-6.
141. Shen, R., A.B. Olshen, and M. Ladanyi, *Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis.* Bioinformatics, 2009. **25**(22): p. 2906-12.
142. Kirk, P., et al., *Bayesian correlated clustering to integrate multiple datasets.* Bioinformatics, 2012. **28**(24): p. 3290-7.
143. Barkai, N. and S. Leibler, *Robustness in simple biochemical networks.* Nature, 1997. **387**(6636): p. 913-7.
144. Albeck, J.G., et al., *Quantitative analysis of pathways controlling extrinsic apoptosis in single cells.* Molecular cell, 2008. **30**(1): p. 11-25.
145. Soltis, A.R. and J.J. Saucerman, *Synergy between CaMKII substrates and beta-adrenergic signaling in regulation of cardiac myocyte Ca(2+) handling.* Biophysical journal, 2010. **99**(7): p. 2038-47.
146. Saucerman, J.J., et al., *Modeling beta-adrenergic control of cardiac myocyte contractility in silico.* The Journal of biological chemistry, 2003. **278**(48): p. 47997-8003.
147. Schoeberl, B., et al., *Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors.* Nature biotechnology, 2002. **20**(4): p. 370-5.
148. Gutenkunst, R.N., et al., *Universally sloppy parameter sensitivities in systems biology models.* PLoS computational biology, 2007. **3**(10): p. 1871-78.
149. Soltis, A.R. and J.J. Saucerman, *Robustness portraits of diverse biological networks conserved despite order-of-magnitude parameter uncertainty.* Bioinformatics, 2011. **27**(20): p. 2888-94.
150. Albert, I., et al., *Boolean network simulations for life scientists.* Source code for biology and medicine, 2008. **3**: p. 16.
151. Morris, M.K., et al., *Training signaling pathway maps to biochemical data with constrained fuzzy logic: quantitative analysis of liver cell responses to inflammatory stimuli.* PLoS computational biology, 2011. **7**(3): p. e1001099.
152. Orth, J.D., I. Thiele, and B.O. Palsson, *What is flux balance analysis?* Nature biotechnology, 2010. **28**(3): p. 245-8.
153. Gudmundsson, S. and I. Thiele, *Computationally efficient flux variability analysis.* BMC bioinformatics, 2010. **11**: p. 489.
154. Yizhak, K., et al., *Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model.* Bioinformatics, 2010. **26**(12): p. i255-60.
155. Huang, S.S. and E. Fraenkel, *Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks.* Science signaling, 2009. **2**(81): p. ra40.
156. Albert, R., *Network inference, analysis, and modeling in systems biology.* The Plant cell, 2007. **19**(11): p. 3327-38.
157. Lander, A.D., *The edges of understanding.* BMC biology, 2010. **8**: p. 40.
158. Tuncbag, N., et al., *Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem.* Journal of computational biology : a journal of computational molecular cell biology, 2013. **20**(2): p. 124-36.
159. Tuncbag, N., et al., *Network-based interpretation of diverse high-throughput datasets through the Omics Integrator software package.* PLoS computational biology, 2016. **(In press)**.
160. Huang, S.S., et al., *Linking proteomic and transcriptional data through the interactome and epigenome reveals a map of oncogene-induced signaling.* PLoS computational biology, 2013. **9**(2): p. e1002887.

161. Lan, A., et al., *ResponseNet: revealing signaling and regulatory networks linking genetic and transcriptomic screening data.* Nucleic Acids Res, 2011. **39**(Web Server issue): p. W424-9.
162. Gosline, S.J., et al., *SAMNet: a network-based approach to integrate multi-dimensional high throughput datasets.* Integr Biol (Camb), 2012. **4**(11): p. 1415-27.
163. Paull, E.O., et al., *Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE).* Bioinformatics, 2013. **29**(21): p. 2757-64.
164. Radonjic, M., et al., *Differential effects of drug interventions and dietary lifestyle in developing type 2 diabetes and complications: a systems biology analysis in LDLr-/- mice.* PloS one, 2013. **8**(2): p. e56122.
165. Kelder, T., et al., *Exploring pathway interactions in insulin resistant mouse liver.* BMC Systems Biology, 2011. **5**(1): p. 127.
166. Li, C.C., et al., *Maternal obesity and diabetes induces latent metabolic defects and widespread epigenetic changes in isogenic mice.* Epigenetics, 2013. **8**(6): p. 602-11.
167. Nilsson, E., et al., *Epigenetic Alterations in Human Liver From Subjects With Type 2 Diabetes in Parallel With Reduced Folate Levels.* The Journal of clinical endocrinology and metabolism, 2015. **100**(11): p. E1491-E1501.
168. Leung, A., et al., *Open chromatin profiling in mice livers reveals unique chromatin variations induced by high fat diet.* The Journal of biological chemistry, 2014. **289**(34): p. 23557-67.
169. Bochkis, I.M., et al., *Genome-wide location analysis reveals distinct transcriptional circuitry by paralogous regulators Foxa1 and Foxa2.* PLoS genetics, 2012. **8**(6): p. e1002770.
170. Shin, D.-J., et al., *Genome-wide analysis of FoxO1 binding in hepatic chromatin: Potential involvement of FoxO1 in linking retinoid signaling to hepatic gluconeogenesis.* Nucleic acids research, 2012. **40**(22): p. 11499-11509.
171. Johnson, M.E., et al., *Genome-Wide Analyses of ChIP-Seq Derived FOXA2 DNA Occupancy in Liver Points to Genetic Networks Underpinning Multiple Complex Traits.* The Journal of clinical endocrinology and metabolism, 2014. **99**(8): p. E1580-E1585.
172. Burdin, D.V., et al., *Diabetes-linked transcription factor HNF4α regulates metabolism of endogenous methylarginines and β-aminoisobutyric acid by controlling expression of alanine-glyoxylate aminotransferase 2.* Scientific Reports, 2016. **6**: p. 35503.
173. Kim, Y.-C., et al., *Liver ChIP-seq analysis in FGF19-treated mice reveals SHP as a global transcriptional partner of SREBP-2.* Genome biology, 2015. **16**(1): p. 268.
174. Deng, W.J., et al., *Proteome, phosphoproteome, and hydroxyproteome of liver mitochondria in diabetic rats at early pathogenic stages.* Molecular & cellular proteomics : MCP, 2010. **9**(1): p. 100-16.
175. Guo, Y., et al., *Quantitative proteomic and functional analysis of liver mitochondria from high fat diet (HFD) diabetic mice.* Molecular & cellular proteomics : MCP, 2013. **12**(12): p. 3744-58.
176. Sabido, E., et al., *Targeted proteomics reveals strain-specific changes in the mouse insulin and central metabolic pathways after a sustained high-fat diet.* Molecular systems biology, 2013. **9**: p. 681.
177. Wu, Y., et al., *Multilayered genetic and omics dissection of mitochondrial activity in a mouse reference population.* Cell, 2014. **158**(6): p. 1415-30.
178. Sas, K.M., et al., *Metabolomics and diabetes: analytical and computational approaches.* Diabetes, 2015. **64**(3): p. 718-32.
179. Kim, H.J., et al., *Metabolomic analysis of livers and serum from high-fat diet induced obese mice.* Journal of proteome research, 2011. **10**(2): p. 722-31.
180. Wang, T.J., et al., *Metabolite profiles and the risk of developing diabetes.* Nature medicine, 2011. **17**(4): p. 448-453.
181. Wang, T.J., et al., *2-Aminoadipic acid is a biomarker for diabetes risk.* The Journal of clinical investigation, 2013. **123**(10): p. 4309-4317.
182. Floegel, A., et al., *Identification of Serum Metabolites Associated With Risk of Type 2 Diabetes Using a Targeted Metabolomic Approach.* Diabetes, 2013. **62**(2): p. 639-648.
183. Nikiforova, V.J., et al., *Glyoxylate, a new marker metabolite of type 2 diabetes.* Journal of diabetes research, 2014. **2014**: p. 685204.
184. Kirpich, I.A., et al., *Integrated hepatic transcriptome and proteome analysis of mice with high-fat diet-induced nonalcoholic fatty liver disease.* The Journal of nutritional biochemistry, 2011. **22**(1): p. 38-45.
185. Meierhofer, D., C. Weidner, and S. Sauer, *Integrative analysis of transcriptomics, proteomics, and metabolomics data of white adipose and liver tissue of high-fat diet and rosiglitazone-treated insulin-resistant mice identified pathway alterations and molecular hubs.* Journal of proteome research, 2014. **13**(12): p. 5592-602.

186. Kamburov, A., et al., *Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA.* Bioinformatics, 2011. **27**(20): p. 2917-8.
187. Szklarczyk, D., et al., *The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored.* Nucleic acids research, 2011. **39**(Database issue): p. D561-8.
188. Patel, V.R., et al., *CircadiOmics: integrating circadian genomics, transcriptomics, proteomics and metabolomics.* Nature methods, 2012. **9**(8): p. 772-3.
189. Varemo, L., I. Nookaew, and J. Nielsen, *Novel insights into obesity and diabetes through genome-scale metabolic modeling.* Frontiers in physiology, 2013. **4**: p. 92.
190. Uhlen, M., et al., *Proteomics. Tissue-based map of the human proteome.* Science, 2015. **347**(6220): p. 1260419.
191. Mardinoglu, A., et al., *Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease.* Nature communications, 2014. **5**: p. 3083.
192. Varemo, L., et al., *Proteome- and transcriptome-driven reconstruction of the human myocyte metabolic network and its use for identification of markers for diabetes.* Cell reports, 2015. **11**(6): p. 921-33.
193. Lee, S., et al., *Integrated Network Analysis Reveals an Association between Plasma Mannose Levels and Insulin Resistance.* Cell metabolism, 2016. **24**(1): p. 172-84.
194. Azimifar, S.B., et al., *Cell-type-resolved quantitative proteomics of murine liver.* Cell metabolism, 2014. **20**(6): p. 1076-87.
195. Eckel-Mahan, K.L., et al., *Reprogramming of the circadian clock by nutritional challenge.* Cell, 2013. **155**(7): p. 1464-78.
196. Wijekoon, E.P., et al., *Amino acid metabolism in the Zucker diabetic fatty rat: effects of insulin resistance and of type 2 diabetes.* Canadian journal of physiology and pharmacology, 2004. **82**(7): p. 506-14.
197. Schrem, H., J. Klempnauer, and J. Borlak, *Liver-enriched transcription factors in liver function and development. Part I: the hepatocyte nuclear factor network and liver-specific gene expression.* Pharmacological reviews, 2002. **54**(1): p. 129-58.
198. Schrem, H., J. Klempnauer, and J. Borlak, *Liver-enriched transcription factors in liver function and development. Part II: the C/EBPs and D site-binding protein in cell cycle control, carcinogenesis, circadian gene regulation, liver regeneration, apoptosis, and liver-specific gene regulation.* Pharmacological reviews, 2004. **56**(2): p. 291-330.
199. Pinero, J., et al., *DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes.* Database : the journal of biological databases and curation, 2015. **2015**: p. bav028.
200. Daimon, M., et al., *Association of the clusterin gene polymorphisms with type 2 diabetes mellitus.* Metabolism: clinical and experimental, 2011. **60**(6): p. 815-22.
201. Kwon, M.J., et al., *Deficiency of clusterin exacerbates high-fat diet-induced insulin resistance in male mice.* Endocrinology, 2014. **155**(6): p. 2089-101.
202. Choe, C.U., et al., *L-arginine:glycine amidinotransferase deficiency protects from metabolic syndrome.* Human molecular genetics, 2013. **22**(1): p. 110-23.
203. Wang, Z., O.J. Shah, and T. Hunter, *The transcriptional coactivators p/CIP and SRC-1 control insulin resistance through IRS1 in obesity models.* PloS one, 2012. **7**(7): p. e36961.
204. Agius, L., *Targeting hepatic glucokinase in type 2 diabetes: weighing the benefits and risks.* Diabetes, 2009. **58**(1): p. 18-20.
205. Caro, J.F., et al., *Liver glucokinase: decreased activity in patients with type II diabetes.* Hormone and metabolic research = Hormon- und Stoffwechselforschung = Hormones et métabolisme, 1995. **27**(1): p. 19-22.
206. Futamura, M., et al., *An allosteric activator of glucokinase impairs the interaction of glucokinase and glucokinase regulatory protein and regulates glucose metabolism.* The Journal of biological chemistry, 2006. **281**(49): p. 37668-74.
207. Lin, Y. and Z. Sun, *Thyroid hormone potentiates insulin signaling and attenuates hyperglycemia and insulin resistance in a mouse model of type 2 diabetes.* British journal of pharmacology, 2011. **162**(3): p. 597-610.
208. Martagon, A.J., et al., *The amelioration of hepatic steatosis by thyroid hormone receptor agonists is insufficient to restore insulin sensitivity in ob/ob mice.* PloS one, 2015. **10**(4): p. e0122987.
209. Kim, G.H., et al., *Regulation of hepatic insulin sensitivity by activating signal co-integrator-2.* The Biochemical journal, 2012. **447**(3): p. 437-47.
210. Mottis, A., L. Mouchiroud, and J. Auwerx, *Emerging roles of the corepressors NCoR1 and SMRT in homeostasis.* Genes & development, 2013. **27**(8): p. 819-35.

211. Mihalik, S.J., et al., *Participation of two members of the very long-chain acyl-CoA synthetase family in bile acid synthesis and recycling*. The Journal of biological chemistry, 2002. **277**(27): p. 24771-9.
212. Johnson, M.R., et al., *Purification and characterization of bile acid-CoA:amino acid N-acyltransferase from human liver.* The Journal of biological chemistry, 1991. **266**(16): p. 10227-33.
213. Garzel, B., et al., *The role of bile salt export pump gene repression in drug-induced cholestatic liver toxicity.* Drug metabolism and disposition: the biological fate of chemicals, 2014. **42**(3): p. 318-22.
214. Ohtsuka, H., et al., *Farnesoid X receptor, hepatocyte nuclear factors 1alpha and 3beta are essential for transcriptional activation of the liver-specific organic anion transporter-2 gene.* Journal of gastroenterology, 2006. **41**(4): p. 369-77.
215. Pyo, J.O., et al., *Essential roles of Atg5 and FADD in autophagic cell death: dissection of autophagic cell death into vacuole formation and cell death.* The Journal of biological chemistry, 2005. **280**(21): p. 20722-9.
216. Wang, J., et al., *Caspase-10 is an initiator caspase in death receptor signaling*. Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(24): p. 13884-8.
217. Lee, S.C. and S. Pervaiz, *Apoptosis in the pathophysiology of diabetes mellitus.* The international journal of biochemistry & cell biology, 2007. **39**(3): p. 497-504.
218. Wang, K., *Molecular mechanisms of hepatic apoptosis.* Cell death & disease, 2014. **5**: p. e996.
219. Williams, A.S., L. Kang, and D.H. Wasserman, *The extracellular matrix and insulin resistance.* Trends in endocrinology and metabolism: TEM, 2015. **26**(7): p. 357-366.
220. Kang, L., et al., *Hyaluronan accumulates with high-fat feeding and contributes to insulin resistance.* Diabetes, 2013. **62**(6): p. 1888-96.
221. Williams, A.S., et al., *Integrin alpha1-null mice exhibit improved fatty liver when fed a high fat diet despite severe hepatic insulin resistance.* The Journal of biological chemistry, 2015. **290**(10): p. 6546-57.
222. Schattenberg, J.M. and M. Schuchmann, *Diabetes and apoptosis: liver.* Apoptosis : an international journal on programmed cell death, 2009. **14**(12): p. 1459-71.
223. Guicciardi, M.E. and G.J. Gores, *Apoptosis: a mechanism of acute and chronic liver injury.* Gut, 2005. **54**(7): p. 1024-33.
224. Sabio, G., et al., *A stress signaling pathway in adipose tissue regulates hepatic insulin resistance.* Science, 2008. **322**(5907): p. 1539-43.
225. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.
226. Hansen, K.D., R.A. Irizarry, and Z. Wu, *Removing technical variability in RNA-seq data using conditional quantile normalization.* Biostatistics, 2012. **13**(2): p. 204-16.
227. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome biology, 2014. **15**(12): p. 550.
228. Langmead, B., *Aligning short sequencing reads with Bowtie.* Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.], 2010. **Chapter 11**: p. Unit 11 7.
229. Zhang, Y., et al., *Model-based analysis of ChIP-Seq (MACS).* Genome biology, 2008. **9**(9): p. R137.
230. Zhou, F., et al., *Genome-scale proteome quantification by DEEP SEQ mass spectrometry.* Nature communications, 2013. **4**: p. 2171.
231. Ficarro, S.B., et al., *Improved electrospray ionization efficiency compensates for diminished chromatographic resolution and enables proteomics analysis of tyrosine signaling in embryonic stem cells.* Analytical chemistry, 2009. **81**(9): p. 3440-7.
232. Parikh, J.R., et al., *multiplierz: an extensible API based desktop environment for proteomics data analysis.* BMC bioinformatics, 2009. **10**: p. 364.
233. Oberg, A.L. and D.W. Mahoney, *Statistical methods for quantitative mass spectrometry proteomic experiments with labeling.* BMC bioinformatics, 2012. **13 Suppl 16**: p. S7.
234. Anders, S. and W. Huber, *Differential expression analysis for sequence count data.* Genome biology, 2010. **11**(10): p. R106.
235. Gordon, D.B., et al., *TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs.* Bioinformatics, 2005. **21**(14): p. 3164-5.
236. Villaveces, J.M., et al., *Merging and scoring molecular interactions utilising existing community standards: tools, use-cases and a case study.* Database : the journal of biological databases and curation, 2015. **2015**.
237. Blondel, V.D., et al., *Fast unfolding of communities in large networks.* Journal of Statistical Mechanics: Theory and Experiment, 2008. **2008**(10): p. P10008-P10008.

238. Jewison, T., et al., *SMPDB 2.0: big improvements to the Small Molecule Pathway Database.* Nucleic acids research, 2014. **42**(Database issue): p. D478-84.
239. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(43): p. 15545-50.
240. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome research, 2003. **13**(11): p. 2498-504.
241. Carpenter, A.E., et al., *CellProfiler: image analysis software for identifying and quantifying cell phenotypes.* Genome biology, 2006. **7**(10): p. R100.
242. Schroeder-Gloeckler, J.M., et al., *CCAAT/enhancer-binding protein beta deletion reduces adiposity, hepatic steatosis, and diabetes in Lepr(db/db) mice.* The Journal of biological chemistry, 2007. **282**(21): p. 15717-29.
243. Kwon, H.S., et al., *Protein kinase B-alpha inhibits human pyruvate dehydrogenase kinase-4 gene induction by dexamethasone through inactivation of FOXO transcription factors.* Diabetes, 2004. **53**(4): p. 899-910.
244. Frey, B.J. and D. Dueck, *Clustering by passing messages between data points.* Science, 2007. **315**(5814): p. 972-6.
245. Freedman, N.J. and R.J. Lefkowitz, *Desensitization of G protein-coupled receptors.* Recent progress in hormone research, 1996. **51**: p. 319-51; discussion 352-3.
246. De Vries, L., et al., *The regulator of G protein signaling family.* Annual review of pharmacology and toxicology, 2000. **40**: p. 235-71.
247. Bansal, G., K.M. Druey, and Z. Xie, *R4 RGS proteins: regulation of G-protein signaling and beyond.* Pharmacology & therapeutics, 2007. **116**(3): p. 473-95.
248. Gold, S.J., et al., *Regulation of RGS proteins by chronic morphine in rat locus coeruleus.* The European journal of neuroscience, 2003. **17**(5): p. 971-80.
249. Ding, J., et al., *RGS4-dependent attenuation of M4 autoreceptor function in striatal cholinergic interneurons following dopamine depletion.* Nature neuroscience, 2006. **9**(6): p. 832-42.
250. Buckholtz, J.W., et al., *Allelic variation in RGS4 impacts functional and structural connectivity in the human brain.* The Journal of neuroscience : the official journal of the Society for Neuroscience, 2007. **27**(7): p. 1584-93.
251. Gu, Z., Q. Jiang, and Z. Yan, *RGS4 modulates serotonin signaling in prefrontal cortex and links to serotonin dysfunction in a rat model of schizophrenia.* Molecular pharmacology, 2007. **71**(4): p. 1030-9.
252. Owen, V.J., et al., *Expression of RGS3, RGS4 and Gi alpha 2 in acutely failing donor hearts and end-stage heart failure.* European heart journal, 2001. **22**(12): p. 1015-20.
253. Postic, C., et al., *Dual roles for glucokinase in glucose homeostasis as determined by liver and pancreatic beta cell-specific gene knock-outs using Cre recombinase.* The Journal of biological chemistry, 1999. **274**(1): p. 305-15.
254. Bruning, J.C., et al., *A muscle-specific insulin receptor knockout exhibits features of the metabolic syndrome of NIDDM without altering glucose tolerance.* Molecular cell, 1998. **2**(5): p. 559-69.
255. Kim, H.J., et al., *Differential effects of interleukin-6 and -10 on skeletal muscle and liver insulin action in vivo.* Diabetes, 2004. **53**(4): p. 1060-7.
256. Vernia, S., et al., *The PPARalpha-FGF21 hormone axis contributes to metabolic regulation by the hepatic JNK signaling pathway.* Cell metabolism, 2014. **20**(3): p. 512-25.
257. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society Series B-Methodological, 1995. **57**(1): p. 289-300.
258. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nature genetics, 2000. **25**(1): p. 25-9.
259. Biddinger, S.B. and C.R. Kahn, *FROM MICE TO MEN: Insights into the Insulin Resistance Syndromes.* Annual Review of Physiology, 2006. **68**(1): p. 123-158.
260. Farrell, G.C. and C.Z. Larter, *Nonalcoholic fatty liver disease: from steatosis to cirrhosis.* Hepatology, 2006. **43**(2 Suppl 1): p. S99-S112.
261. Tilg, H. and A.R. Moschen, *Insulin resistance, inflammation, and non-alcoholic fatty liver disease.* Trends Endocrinol Metab, 2008. **19**(10): p. 371-9.
262. Bishop, N.A. and L. Guarente, *Two neurons mediate diet-restriction-induced longevity in C. elegans.* Nature, 2007. **447**(7144): p. 545-9.

263. Colman, R.J., et al., *Caloric restriction delays disease onset and mortality in rhesus monkeys.* Science, 2009. **325**(5937): p. 201-4.
264. Yu, H., W. Jia, and Z. Guo, *Reducing Liver Fat by Low Carbohydrate Caloric Restriction Targets Hepatic Glucose Production in Non-Diabetic Obese Adults with Non-Alcoholic Fatty Liver Disease.* Journal of clinical medicine, 2014. **3**(3): p. 1050-63.
265. Larson-Meyer, D.E., et al., *Effect of 6-month calorie restriction and exercise on serum and liver lipids and markers of liver function.* Obesity, 2008. **16**(6): p. 1355-62.
266. Cao, S.X., et al., *Genomic profiling of short- and long-term caloric restriction effects in the liver of aging mice.* Proceedings of the National Academy of Sciences of the United States of America, 2001. **98**(19): p. 10630-5.
267. Cai, Y., et al., *The role of hepatocyte RXR alpha in xenobiotic-sensing nuclear receptor-mediated pathways.* European journal of pharmaceutical sciences : official journal of the European Federation for Pharmaceutical Sciences, 2002. **15**(1): p. 89-96.
268. Wei, P., et al., *The nuclear receptor CAR mediates specific xenobiotic induction of drug metabolism.* Nature, 2000. **407**(6806): p. 920-923.
269. Schmitz, G. and T. Langmann, *Transcriptional regulatory networks in lipid metabolism control ABCA1 expression.* Biochimica et biophysica acta, 2005. **1735**(1): p. 1-19.
270. Braissant, O., et al., *Differential expression of peroxisome proliferator-activated receptors (PPARs): tissue distribution of PPAR-alpha, -beta, and -gamma in the adult rat.* Endocrinology, 1996. **137**(1): p. 354-66.
271. Germain, P., et al., *International Union of Pharmacology. LXIII. Retinoid X receptors.* Pharmacol Rev, 2006. **58**(4): p. 760-72.
272. Rakhshandehroo, M., et al., *Peroxisome proliferator-activated receptor alpha target genes.* PPAR Research, 2010. **2010**.
273. Gervois, P., et al., *Regulation of lipid and lipoprotein metabolism by PPAR activators.* Clinical chemistry and laboratory medicine, 2000. **38**(1): p. 3-11.
274. Koo, S.-H., et al., *The CREB coactivator TORC2 is a key regulator of fasting glucose metabolism.* Nature, 2005. **437**(7062): p. 1109-11.
275. Kane, C.D., et al., *Molecular characterization of novel and selective peroxisome proliferator-activated receptor alpha agonists with robust hypolipidemic activity in vivo.* Mol Pharmacol, 2009. **75**(2): p. 296-306.
276. Mandard, S., M. Muller, and S. Kersten, *Peroxisome proliferator-activated receptor alpha target genes.* Cell Mol Life Sci, 2004. **61**(4): p. 393-416.
277. Peeters, A. and M. Baes, *Role of PPARalpha in Hepatic Carbohydrate Metabolism.* PPAR Res, 2010. **2010**.
278. Barak, Y. and S. Kim, *Genetic manipulations of PPARs: effects on obesity and metabolic disease.* PPAR Res, 2007. **2007**: p. 12781.
279. Boergesen, M., et al., *Genome-Wide Profiling of Liver X Receptor, Retinoid X Receptor, and Peroxisome Proliferator-Activated Receptor α in Mouse Liver Reveals Extensive Sharing of Binding Sites.* Molecular and Cellular Biology, 2012. **32**(4): p. 852-867.
280. Dowman, J.K., J.W. Tomlinson, and P.N. Newsome, *Pathogenesis of non-alcoholic fatty liver disease.* QJM, 2010. **103**(2): p. 71-83.
281. Bruss, M.D., et al., *Calorie restriction increases fatty acid synthesis and whole body fat oxidation rates.* Am J Physiol Endocrinol Metab, 2010. **298**(1): p. E108-16.
282. Hotamisligil, G.S., *Inflammation and metabolic disorders.* Nature, 2006. **444**(7121): p. 860-867.
283. Stanton, M.C., et al., *Inflammatory Signals shift from adipose to liver during high fat feeding and influence the development of steatohepatitis in mice.* J Inflamm (Lond), 2011. **8**: p. 8.
284. Atherton, H.J., et al., *A combined 1H-NMR spectroscopy- and mass spectrometry-based metabolomic study of the PPAR-alpha null mutant mouse defines profound systemic changes in metabolism linked to the metabolic syndrome.* Physiol Genomics, 2006. **27**(2): p. 178-86.
285. Chakravarthy, M.V., et al., *"New" hepatic fat activates PPARalpha to maintain glucose, lipid, and cholesterol homeostasis.* Cell Metab, 2005. **1**(5): p. 309-22.
286. Pineda Torra, I., et al., *Characterization of the human PPARalpha promoter: identification of a functional nuclear receptor response element.* Mol Endocrinol, 2002. **16**(5): p. 1013-28.
287. Turturro, A., et al., *Growth curves and survival characteristics of the animals used in the Biomarkers of Aging Program.* The journals of gerontology. Series A, Biological sciences and medical sciences, 1999. **54**(11): p. B492-501.

288. Smyth, G.K., *Linear models and empirical bayes methods for assessing differential expression in microarray experiments.* Stat Appl Genet Mol Biol, 2004. **3**: p. Article3.
289. Ling, G., et al., *Unbiased, genome-wide in vivo mapping of transcriptional regulatory elements reveals sex differences in chromatin structure associated with sex-specific liver gene expression.* Mol Cell Biol, 2010. **30**(23): p. 5531-44.
290. Guo, Y., et al., *Discovering homotypic binding events at high spatial resolution.* Bioinformatics, 2010. **26**(24): p. 3028-34.
291. Karlic, R., et al., *Histone modification levels are predictive for gene expression.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(7): p. 2926-31.
292. Roider, H.G., et al., *CpG-depleted promoters harbor tissue-specific transcription factor binding signals--implications for motif overrepresentation analyses.* Nucleic Acids Res, 2009. **37**(19): p. 6305-15.
293. Wingender, E., et al., *TRANSFAC: a database on transcription factors and their DNA binding sites.* Nucleic Acids Res, 1996. **24**(1): p. 238-41.
294. MacIsaac, K.D., et al., *A quantitative model of transcriptional regulation reveals the influence of binding location on expression.* PLoS Comput Biol, 2010. **6**(4): p. e1000773.
295. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome biology, 2009. **10**(3): p. R25.
296. Zhang, Y., et al., *Model-based analysis of ChIP-Seq (MACS).* Genome Biol, 2008. **9**(9): p. R137.
297. Seglen, P.O., *Preparation of isolated rat liver cells.* Methods Cell Biol, 1976. **13**: p. 29-83.
298. Vernia, S., et al., *The PPARalpha-FGF21 hormone axis contributes to metabolic regulation by the hepatic JNK signaling pathway.* Cell Metab, 2014. **20**(3): p. 512-25.
299. Bartel, D.P., *MicroRNAs: target recognition and regulatory functions.* Cell, 2009. **136**(2): p. 215-33.
300. Ha, M. and V.N. Kim, *Regulation of microRNA biogenesis.* Nature reviews. Molecular cell biology, 2014. **15**(8): p. 509-24.
301. Friedman, R.C., et al., *Most mammalian mRNAs are conserved targets of microRNAs.* Genome research, 2009. **19**(1): p. 92-105.
302. Selbach, M., et al., *Widespread changes in protein synthesis induced by microRNAs.* Nature, 2008. **455**(7209): p. 58-63.
303. Guo, H., et al., *Mammalian microRNAs predominantly act to decrease target mRNA levels.* Nature, 2010. **466**(7308): p. 835-40.
304. Reinhart, B.J., et al., *The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans.* Nature, 2000. **403**(6772): p. 901-6.
305. Joglekar, M.V., et al., *MicroRNA profiling of developing and regenerating pancreas reveal post-transcriptional regulation of neurogenin3.* Developmental biology, 2007. **311**(2): p. 603-12.
306. Chen, Q., et al., *MiR-34a regulates apoptosis in liver cells by targeting the KLF4 gene.* Cellular & molecular biology letters, 2014. **19**(1): p. 52-64.
307. Zhang, W., J. Liu, and G. Wang, *The role of microRNAs in human breast cancer progression.* Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine, 2014. **35**(7): p. 6235-44.
308. Ikeda, S., et al., *Altered microRNA expression in human heart disease.* Physiological genomics, 2007. **31**(3): p. 367-73.
309. Nelson, P.T., W.X. Wang, and B.W. Rajeev, *MicroRNAs (miRNAs) in neurodegenerative diseases.* Brain pathology, 2008. **18**(1): p. 130-8.
310. Pauley, K.M., S. Cha, and E.K. Chan, *MicroRNA in autoimmunity and autoimmune diseases.* Journal of autoimmunity, 2009. **32**(3-4): p. 189-94.
311. Rottiers, V. and A.M. Naar, *MicroRNAs in metabolism and metabolic disorders.* Nature reviews. Molecular cell biology, 2012. **13**(4): p. 239-50.
312. Williams, M.D. and G.M. Mitchell, *MicroRNAs in insulin resistance and obesity.* Experimental diabetes research, 2012. **2012**: p. 484696.
313. Esau, C., et al., *miR-122 regulation of lipid metabolism revealed by in vivo antisense targeting.* Cell metabolism, 2006. **3**(2): p. 87-98.
314. Kornfeld, J.W., et al., *Obesity-induced overexpression of miR-802 impairs glucose metabolism through silencing of Hnf1b.* Nature, 2013. **494**(7435): p. 111-5.
315. Trajkovski, M., et al., *MicroRNAs 103 and 107 regulate insulin sensitivity.* Nature, 2011. **474**(7353): p. 649-53.

316. Rayner, K.J., et al., *MiR-33 contributes to the regulation of cholesterol homeostasis.* Science, 2010. **328**(5985): p. 1570-3.
317. Jordan, S.D., et al., *Obesity-induced overexpression of miRNA-143 inhibits insulin-stimulated AKT activation and impairs glucose metabolism.* Nature cell biology, 2011. **13**(4): p. 434-46.
318. Zhou, B., et al., *Downregulation of miR-181a upregulates sirtuin-1 (SIRT1) and improves hepatic insulin sensitivity.* Diabetologia, 2012. **55**(7): p. 2032-43.
319. Garcia, D.M., et al., *Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs.* Nature structural & molecular biology, 2011. **18**(10): p. 1139-46.
320. Haecker, I., et al., *Ago HITS-CLIP expands understanding of Kaposi's sarcoma-associated herpesvirus miRNA function in primary effusion lymphomas.* PLoS pathogens, 2012. **8**(8): p. e1002884.
321. Helwak, A., et al., *Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding.* Cell, 2013. **153**(3): p. 654-65.
322. Maragkakis, M., et al., *DIANA-microT web server: elucidating microRNA functions through target prediction.* Nucleic acids research, 2009. **37**(Web Server issue): p. W273-6.
323. Kertesz, M., et al., *The role of site accessibility in microRNA target recognition.* Nature genetics, 2007. **39**(10): p. 1278-84.
324. Gosline, S.J., et al., *Elucidating MicroRNA Regulatory Networks Using Transcriptional, Post-transcriptional, and Histone Modification Measurements.* Cell reports, 2016. **14**(2): p. 310-9.
325. Ruan, Q., et al., *The microRNA-21-PDCD4 axis prevents type 1 diabetes by blocking pancreatic beta cell death.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(29): p. 12030-5.
326. Dey, N., et al., *MicroRNA-21 orchestrates high glucose-induced signals to TOR complex 1, resulting in renal cell pathology in diabetes.* The Journal of biological chemistry, 2011. **286**(29): p. 25586-603.
327. Zhong, X., et al., *miR-21 is a key therapeutic target for renal injury in a mouse model of type 2 diabetes.* Diabetologia, 2013. **56**(3): p. 663-74.
328. Wang, L., et al., *MicroRNAs 185, 96, and 223 repress selective high-density lipoprotein cholesterol uptake through posttranscriptional inhibition.* Molecular and cellular biology, 2013. **33**(10): p. 1956-64.
329. Wang, X.C., et al., *MicroRNA-185 regulates expression of lipid metabolism genes and improves insulin sensitivity in mice with non-alcoholic fatty liver disease.* World journal of gastroenterology, 2014. **20**(47): p. 17914-23.
330. de Guia, R.M., et al., *microRNA-379 couples glucocorticoid hormones to dysfunctional lipid homeostasis.* The EMBO journal, 2015. **34**(3): p. 344-60.
331. Yamada, H., et al., *Associations between circulating microRNAs (miR-21, miR-34a, miR-122 and miR-451) and non-alcoholic fatty liver.* Clinica chimica acta; international journal of clinical chemistry, 2013. **424**: p. 99-103.
332. Oda, Y., et al., *Retinoid X receptor alpha in human liver is regulated by miR-34a.* Biochemical pharmacology, 2014. **90**(2): p. 179-87.
333. Lee, J., et al., *A pathway involving farnesoid X receptor and small heterodimer partner positively regulates hepatic sirtuin 1 levels via microRNA-34a inhibition.* The Journal of biological chemistry, 2010. **285**(17): p. 12604-11.
334. Wu, L., et al., *Profiling peripheral microRNAs in obesity and type 2 diabetes mellitus.* APMIS : acta pathologica, microbiologica, et immunologica Scandinavica, 2015. **123**(7): p. 580-5.
335. Miao, C.G., et al., *MicroRNA-152 modulates the canonical Wnt pathway activation by targeting DNA methyltransferase 1 in arthritic rat model.* Biochimie, 2014. **106**: p. 149-56.
336. Liu, W., et al., *Hepatic miR-378 targets p110alpha and controls glucose and lipid homeostasis by modulating hepatic insulin signalling.* Nature communications, 2014. **5**: p. 5684.
337. Babiarz, J.E., et al., *A role for noncanonical microRNAs in the mammalian brain revealed by phenotypic differences in Dgcr8 versus Dicer1 knockouts and small RNA sequencing.* RNA, 2011. **17**(8): p. 1489-501.
338. Welch, C., Y. Chen, and R.L. Stallings, *MicroRNA-34a functions as a potential tumor suppressor by inducing apoptosis in neuroblastoma cells.* Oncogene, 2007. **26**(34): p. 5017-5022.
339. Allen-Jennings, A.E., et al., *The roles of ATF3 in liver dysfunction and the regulation of phosphoenolpyruvate carboxykinase gene expression.* The Journal of biological chemistry, 2002. **277**(22): p. 20020-5.
340. Jazdzewski, K., et al., *Thyroid hormone receptor beta (THRB) is a major target gene for microRNAs deregulated in papillary thyroid carcinoma (PTC).* The Journal of clinical endocrinology and metabolism, 2011. **96**(3): p. E546-53.

341. Guttilla, I.K. and B.A. White, *Coordinate regulation of FOXO1 by miR-27a, miR-96, and miR-182 in breast cancer cells.* The Journal of biological chemistry, 2009. **284**(35): p. 23204-16.
342. Stittrich, A.-B., et al., *The microRNA miR-182 is induced by IL-2 and promotes clonal expansion of activated helper T lymphocytes.* Nat Immunol, 2010. **11**(11): p. 1057-1062.
343. Zhao, Z., L. Qin, and S. Li, *miR-411 contributes the cell proliferation of lung cancer by targeting FOXO1.* Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine, 2016. **37**(4): p. 5551-60.
344. Gerstein, M.B., et al., *Architecture of the human regulatory network derived from ENCODE data.* Nature, 2012. **489**(7414): p. 91-100.
345. Gurtan, A.M. and P.A. Sharp, *The role of miRNAs in regulating gene expression networks.* Journal of molecular biology, 2013. **425**(19): p. 3582-600.
346. Khorshid, M., et al., *A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets.* Nat Meth, 2013. **10**(3): p. 253-255.
347. Agarwal, V., et al., *Predicting effective microRNA target sites in mammalian mRNAs.* eLife, 2015. **4**.
348. Mukherji, S., et al., *MicroRNAs can generate thresholds in target gene expression.* Nature genetics, 2011. **43**(9): p. 854-859.
349. Nandi, A., et al., *miRNA-regulated dynamics in circadian oscillator models.* BMC systems biology, 2009. **3**(1): p. 45.
350. Nikolov, S., et al., *DYNAMICS OF A miRNA MODEL WITH TWO DELAYS.* Biotechnology & Biotechnological Equipment, 2012. **26**(5): p. 3315-3320.
351. Griffiths-Jones, S., et al., *miRBase: microRNA sequences, targets and gene nomenclature.* Nucleic acids research, 2006. **34**(Database issue): p. D140-4.
352. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.
353. Petersen, K.F. and G.I. Shulman, *Etiology of insulin resistance.* The American journal of medicine, 2006. **119**(5 Suppl 1): p. S10-6.
354. Rangarajan, A. and R.A. Weinberg, *Opinion: Comparative biology of mouse versus human cells: modelling human cancer in mice.* Nature reviews. Cancer, 2003. **3**(12): p. 952-9.
355. Kuhn, A., et al., *Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain.* Nature methods, 2011. **8**(11): p. 945-7.
356. Newman, A.M., et al., *Robust enumeration of cell subsets from tissue expression profiles.* Nature methods, 2015. **12**(5): p. 453-7.
357. Frese, C.K., et al., *Improved peptide identification by targeted fragmentation using CID, HCD and ETD on an LTQ-Orbitrap Velos.* Journal of proteome research, 2011. **10**(5): p. 2377-88.
358. Wei, W., et al., *Deep Coverage Proteomics Identifies More Low-Abundance Missing Proteins in Human Testis Tissue with Q-Exactive HF Mass Spectrometer.* Journal of proteome research, 2016. **15**(11): p. 3988-3997.
359. Singh, A., A. Jayaraman, and J. Hahn, *Modeling regulatory mechanisms in IL-6 signal transduction in hepatocytes.* Biotechnology and bioengineering, 2006. **95**(5): p. 850-62.
360. Hill, S.M., et al., *Inferring causal molecular networks: empirical assessment through a community-based effort.* Nat Meth, 2016. **13**(4): p. 310-318.
361. Hubner, N.C., et al., *Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions.* The Journal of cell biology, 2010. **189**(4): p. 739-54.
362. Hein, M.Y., et al., *A human interactome in three quantitative dimensions organized by stoichiometries and abundances.* Cell, 2015. **163**(3): p. 712-23.
363. Hartstra, A.V., et al., *Insights Into the Role of the Microbiome in Obesity and Type 2 Diabetes.* Diabetes care, 2015. **38**(1): p. 159-165.