CHARACTERIZING AND ANALYZING DISEASE-RELATED OMICS DATA USING NETWORK MODELING APPROACHES

BY

CHUNJING WANG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Chemical Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Doctoral Committee:

Associate Professor Nathan D. Price, Chair
Professor Deborah Leckband
Professor Jian Ma
Associate Professor Christopher Rao

ABSTRACT

Systems biology explores how the components that constitute a biological system interact with each other to produce biological phenotypes. A number of tools for comprehensive and high-throughput measurements of DNA/RNA, protein and metabolites have been developed. Each of these technologies helps to characterize individual components of the genome, proteome or metabolome and offers a distinct perspective about the system structure. My dissertation aims to characterize and analyze multiple types of omics data using existing and novel network-based approaches to better understand disease development mechanisms and improve disease diagnosis and prognosis.

The transcriptome reflects the expression level of mRNAs in single cells or a population of cells. Understanding the transcriptome is an essential part of understanding organism development and disease. The first part of my thesis work focused on analyzing transcriptome data to characterize aggressiveness and heterogeneity of human astrocytoma, the most common glioma with a strikingly high mortality rate. A large-scale global gene expression analysis was performed to analyze gene expression profiles representing hundreds of samples generated by oligonucleotide microarrays. I employed a combination of gene- and network-based approaches to investigate the genetic and biological mechanisms implicated in observed phenotypic differences. I observed increasing dysregulation with increasing tumor grade and concluded that *transcriptomic heterogeneity*, observed at the population scale, is generally correlated with increasingly aggressive phenotypes. Heterogeneity in high-grade astrocytomas also manifests as differences in clinical outcomes and significant efforts had been devoted to identify subtypes within high-grade astrocytomas that have large differences in prognosis. I developed an automated network screening approach which could identify networks capable of predicting subtypes with differential survival in high-grade astrocytomas.

The proteome represents the translated product of the mRNA, and proteomics measurement provides a direct estimate of protein abundance. For the second part of my Ph.D. research, I analyzed mouse brain protein measurements collected by the iTRAQ technology to query and identify dynamically perturbed modules in progressive mouse models of glioblastoma. Network

behavior changes in early, middle and late stages of tumor development in genetically engineered mouse were tracked and 19 genes were selected for further confirmation of their roles in glioblastoma progression. In addition to this specific application to mouse glioblastoma data, the general pipeline represented a novel effort to isolate pathway-level responses to perturbations (e.g., brain tumor formation and progression) from large-scale proteomics data and could be applied in analyzing proteomics data from a variety of different contexts.

The metabolome reflects biological information related to biochemical processes and metabolic networks involving metabolites. Metabolomics data can give an instantaneous snapshot of the current state of the cell and thus offers a distinct view of the effects of diet, drugs and disease on the model organism. The third part of my thesis is dedicated to building and refining genome-scale *in silico* metabolic models for mouse, in order to investigate how the metabolic model responds differently under different conditions (e.g., diabetic vs. normal). This project was completed in two stages: first, I examined the state-of-art genome-scale mouse metabolic model, identified its limitations, and then improved and refined its functionality; second, I created the first liver-specific metabolic models from the generic mouse models by pruning reactions that lack genetic evidence of presence, and then adding liver-specific reactions that represent the characteristics and functions of the mouse liver. Finally, I reconstructed two liver metabolic models for mouse, with one for the normal (control) strain and one for mouse diabetic strains. These two models were compared physiologically to infer metabolic genes that were most impacted by the onset of diabetes.

ACKNOWLEDGEMENT

When I first joined Nathan's research group in 2008, I had no much prior knowledge in systems biology, or even research in general: my humble research experience back in college did not prove helpful or relevant. When I was talking to various professors with their work and research interest and trying to find a group fitted me, I asked Nathan if he liked systems biology; Nathan answered by saying: "some love when they do, some do what they love." In the following years, I kept remembering these words, especially during the down times, when I was not sure if the work I was doing was actually leading anywhere. It taught me the value of commitment and responsibility and it will continue to influence me in my future life, as I am preparing to start my career. Nathan is the most encouraging and positive person I have ever met. I like his attitude to life, to research and I enjoyed a lot of conversations with him. He also gave me a lot of career advices, since I stepped into the job market, not to mention the recommendations and references he wrote for me.

Nathan's group was young and small when I first joined in 2008, and I was his first student coming from China. I am happy to see it has grown into a large, mature and diversified group full of experts from all areas of systems biology and from all parts of the world. I am fortunate to meet a lot of excellent researchers in this research group. Through numerous stimulating discussions with them, I formed new ideas, learned new approaches and knowledge, developed a positive attitude and confidence to overcome temporary difficulties. I would like to thank James Eddy, and Cory Funk for their input and efforts on one of my research projects. I also would like to thank Julie Bletz, Caroline Milne and Yuliang Wang, for their insightful and valuable suggestions on many documents I have written, including this dissertation. Even though I am a

foreign student seeking education by myself in a foreign country, I never feel alone; even though my English is far from perfect, I never feel language set me apart from my colleagues. I could always find a solution, or at least suggestions to a problem, no matter it is from life or from research. I want to mention a few special names: Yuliang Wang, Shuyi Ma, Sriram Chandrasekaran, Areejit Samal, Jaeyun Sung and Matthew Benedict. They have become great friends of mine, and I expect the friendship will expand far beyond the graduate study.

I also would like to thank the remaining members of my thesis committee, Dr. Deborah Leckband, Dr. Jian Ma, and Dr. Christopher Rao. I appreciate their feedback during my preliminary exam and the questions they asked, the perspectives they offered, which made me question things I had been taking for granted. As I buckle up for my defense exam, I hope I could rise to their expectations and answer their questions with quality and well-thought answers, which reflect my effort and knowledge in the past five years.

Last but not least, I would like to thank my family. My parents were never in favor of me going for a doctorate degree, but as I really started my graduate life in Illinois, they provided all kinds of support: emotionally, academically and financially. I am also extremely grateful to my twin sister, who has been my best friend since we were born. I am grateful to the companionship, the emotional support, and constant encouragement and trust she provided. She is always ready to give a patient ear to hear what I want to say, and to say what I want to hear.

I could not possibly name everyone who left a mark in my life in the past few years. They appeared in my life for a reason: they taught me lessons, they made me ponder deeply and they influenced me in a positive way. Altogether they made me become a better person. A lot of people had disappeared from my life, but I will be always grateful for the things they taught me.

I will make sure to give back when I am capable of helping others, to make the happy, helpful and friendly cells to replicate and spread, amidst the cruel and aggressive cancer cells we research on.

Table of Contents

CHAPTER 1: INTRODUCTION AND OVERVIEW

1.1 A SYSTEMS APPROACH TO EXPLORING LIVING ORGANISMS

More than a decade has passed since the term systems biology has been introduced into the language of modern biology [1]. Over the years, its definition has expanded greatly in its width and depth, but one central aspect of systems biology remains unique: it studies how the components that constitute the biological system interact with each other. Disease is becoming more generally perceived as the result of one or more genetically or environmentally perturbed biological networks [2]. As such, to better understand genotype-to-phenotype relationships, we must focus our attention more on the interaction and dynamics of biological *systems* instead of only looking at the individual components of biological processes. This systems approach is in contrast to the classical reductionist approach, where biological systems are dissected into their constituent components. This traditional approach was once successful in the early days of biology, but it has reached its limits after more and more biologists realized the complexities of living systems cannot be explained fully by studying and viewing the components as disparate or disconnected. The system-level perspective offers an alternative to explain how individual pieces create the whole, or how genes interact to create a system-wide phenotype and behavior [3]. The different emphases of the two views are illustrated in **Figure 1.1**.

In order to learn and understand how fundamental biological processes interact with each other, we need to connect two types of information together: the DNA sequence of the genome and the environmental signals and information that operate through living organisms to generate phenotypic responses [4]. To follow the dynamic networks and learn their responses to perturbations, a number of tools aiming for comprehensive and high-throughput measurements of DNA/RNA, protein and metabolites have been developed. Each of them helps to characterize individual components of the genome, proteome or metabolome and offers a different slice of information about the model organism. In the next section, I will give a brief overview of the different high-throughput technologies to address specific hypothesis-driven questions.

1.2 OMICS DATA TO PROFILE AND CHARACTERIZE DISEASE

The first type of omics data is transcriptomics. The *transcriptome* is a collective term for different RNA molecules, including messenger RNA (mRNA), which carries genetic information transcribed from DNA; ribosomal RNA (rRNA), the RNA component of the ribosome that is essential for protein synthesis; transfer RNA (tRNA) which physically connects nucleic acids and amino acids, as well as non-coding RNAs (RNAs that are not translated into proteins). The study of transcriptomics, also known as expression profiling, quantifies the expression level of mRNAs in a given cell population [5]. Understanding the transcriptome is essential to understanding development and disease: by measuring the differentially expressed transcripts under different conditions (e.g. disease vs. control), we could infer the functional elements of the genome and reveal key players in disease initiation and development. Different high-throughput technologies have been developed to measure the transcriptome, including the hybridization-based microarray as well as the newer, emerging sequence-based technology called RNA-seq. The core biological principle behind microarray is the hybridization between two complementary DNA strands, which pair up with each other by forming hydrogen bonds (**Figure 1.2a**). In contrast to microarray methods, RNA sequencing technologies directly measure the cDNA sequence and are able to provide a more concise estimate of the gene expression value [5]. However, sample collection using microarray hybridization is more affordable, more accessible and many more samples are available in data repositories, compared to data collected by RNA sequencing. A significant portion of my thesis was devoted to studying microarray data of brain cancer patients: collecting, normalizing, and comparing gene expression levels under different conditions, and developing data processing pipelines to better characterize brain cancer initialization and progression.

The transcriptome can be seen as a precursor for the *proteome*, which represents the translated product of the mRNA. However, mRNA is not always translated into protein [6] and mRNA level is not the sole factor determining translated protein content. What further complicates the picture is the fact that proteins may undergo a wide variety of chemical modifications after translation, collectively known as *post-translational modification*.

Alternative splicing of the transcripts, where a single gene or transcript encode for multiple proteins [7] also explain why mRNA is often found not to be correlated highly with protein content [8]. Proteomics confirms the presence of protein, provides a direct measure of the amount of protein present, and gives a better understanding of the state of the living organism than genomics [9].

Mass spectrometry (mass spec) is an important technology to characterize protein content. It can be used as a valuable tool to identify and probe the covalent structure of proteins [10]. Another use of mass spec in proteomics is protein quantification. iTRAQ is a non-gel-based technique developed to quantify proteins by analyzing the derivatization of primary amino groups in proteins using isobaric tags [11]. Specifically, iTRAQ facilitates the comparative analysis of peptides and proteins in different conditions. The proteomics study in my Ph.D. research analyzed mouse brain protein measurements collected by the iTRAQ method and developed a general pipeline to isolate pathway-level responses to perturbations (e.g. brain tumor formation and progression) from large-scale proteomics data.

In addition to transcriptomics and proteomics, metabolomics is a systems biology view on the interactions of metabolic pathways within an organics, which offers us fresh insights into the effects of diet, drugs, and disease. The *metabolome* reflects biological information exists in biochemical processes involving metabolites, which are the intermediates and end products of metabolism [12]. While transcriptomic and proteomic analyses do not tell the complete story of the underlying processes in a cell, metabolomics study can give an instantaneous snapshot of the current state of the cell.

After the omics data have been generated, the next step is to identify all the components of a system, establishing their interactions and assessing their dynamics [13]. An effective way to integrate and interrogate these disparate types of information is using *in silico* models. *In silico* or computational modeling lies at the core of systems biology and it has evolved to be a valuable, fast and accurate tool to predict responses to various hypotheses. With increasing omics and clinical data, researchers have built complete and more realistic models that are beginning to produce lab-proven results. These dry lab models help to frame more focused questions and design better laboratory experiments and clinical trial protocols. One type of computational model is a *genome-scale metabolic model*. A metabolic model is a

mathematical representation of the biochemical transformations in the metabolic network of an organism. A network consists of multiple reactions happening in order, and a reaction consists of reactants and products (metabolites) as well as the genes catalyzing the reaction (**Figure 1.3**). If whole genome sequences are integrated with metabolic biochemical networks, a *genome-scale metabolic model* is constructed. The third part of my thesis is dedicated to building and refining genome-scale *in silico* metabolic models for mouse in order to study how the metabolic model respond differently under different conditions (e.g. diabetic vs. normal).

1.3 DISSERTATION ORGANIZATION

The goals of the work presented in my dissertation were to 1) investigate and explore cancer aggressiveness and heterogeneity in the context of human astrocytoma, using *transcriptomic data* 2) develop a framework to learn and quantify network-level changes in response to tumor progression in the context of mouse glioblastoma, using *proteomics data* 3) develop and refine genome-wide *metabolic models* for mouse liver, under normal and diabetic conditions, using the mouse genome-scale metabolic network reconstruction as a starting point.

In virtually all of my research projects, I was fortunate enough to interact and collaborate with many biologists, computer scientists and bioinformatics scientists. Without the numerous inspiring discussions and intellectual exchanges of information, none of the projects could have reached this far. In subsequent chapters of this dissertation, I have used the singular "I" to clarify ideas and analysis that I was directly responsible and "we" to indicate a more collaborative effort.

The chapters in this dissertation are organized as follows:

Chapter 1: Introduces the concept of systems biology and how transcriptomic, proteomic and metabolomic data each offer a distinct systems perspective to study the living organism.

Chapter 2: Describes and summarizes the major observations and key results in the study of human astrocytoma using large-scale gene expression profiles.

Chapter 3: Presents the discovery of a novel prognostic network that could possibly distinguish different subtypes within aggressive human gliomas.

Chapter 4: Describes a computational framework to discover key players in genetically engineered mice with induced mutations to drive glioma progression.

Chapter 5: Examines the generic mouse model and suggests key functionality improvements that will provide a foundation on which to build metabolic models for mouse liver. Subsequently, I present the pipeline to reconstruct a liver model from the generic mouse model, and explain the construction processes of two versions: normal and diabetic liver models.

Chapter 6: Provides a summary and conclusions for the work presented, and list future directions of my work.

1.4 CHAPTER 1 FIGURES

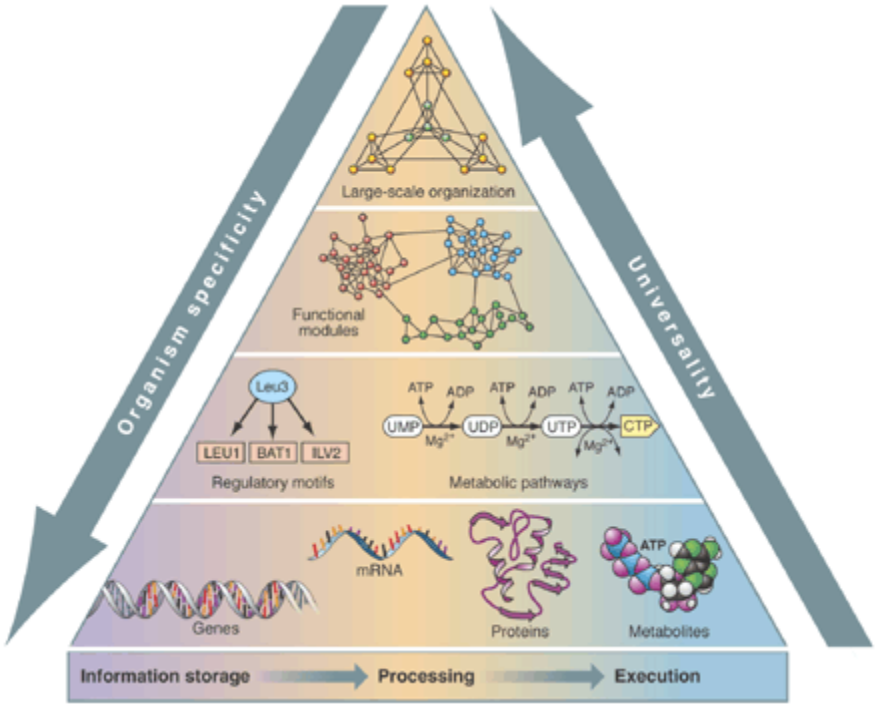


Figure 1.1 The two views on biological systems: the reductionist view knows the precise states of all the organs and cells in the body; the systems view takes account of interactions among all components, and offers us a global perspective on the system as a whole. Adapted from [14] **.**

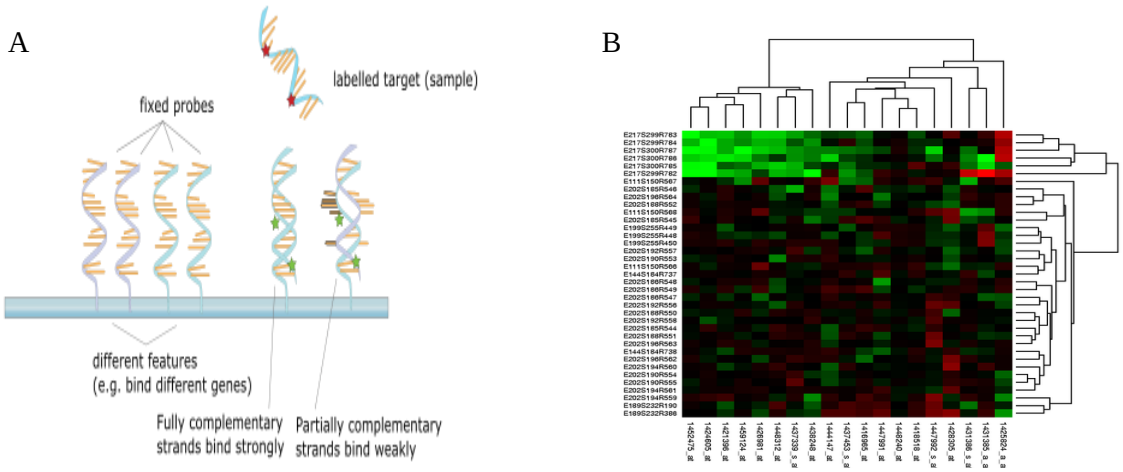


Figure 1.2 A) Hybridization step in microarray B) Heapmap of gene expression values

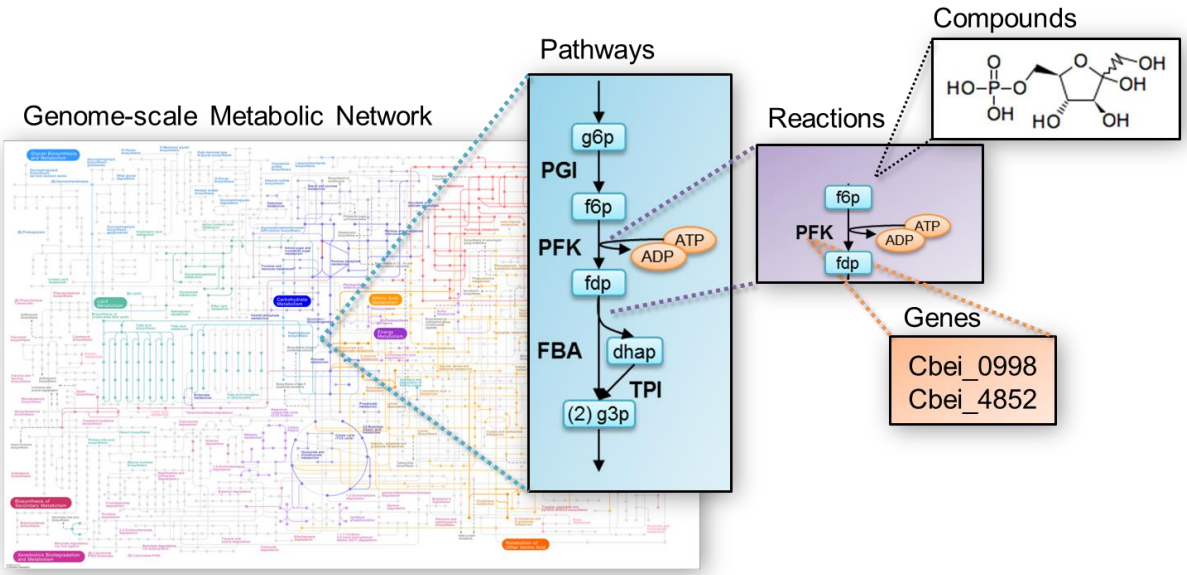


Figure 1.3 Structure of genome-scale metabolic network. Adapted from Systems Biology lecture slides, B.Ø. Palsson, UCSD

CHAPTER 2 UTILIZING TRANSCRIPTOMIC PROFILING TO EXPLORE AGGRESSIVENESS AND HETEROGENEITY OF BRAIN CANCER[1]

2.1 INTRODUCTION TO HUMAN ASTROCYTOMA

Primary brain tumors comprise less than 2% of all human cancers but have strikingly high mortality rates. Glioma, the most prevalent primary brain tumor, accounts for ~42% of all adult brain tumors [15]. The most common gliomas, in turn, are astrocytomas, believed to originate from astrocytes [16, 17]. Astrocytomas are classified from grade 1 (least aggressive) to grade 4 (most aggressive) based on the World Health Organization (WHO) grading system [18].

Presented here is my work on the analysis of the different grades of astrocytoma (excluding pilocytic astrocytoma, with normal brain tissues taken as control) to identify both distinct and common molecular states across grades. I have employed a combination of gene- and network-based approaches (**Figure 2.1**) to investigate the genetic and biological mechanisms implicated in observed phenotypic differences.

Grade 1 tumors (pilocytic astrocytomas) represent distinct pathological and biological entities compared with other tumors [19] and thus were not included in this study. As such, I henceforth considered only grades 2 through 4. Grade 2 (G2) and grade 3 (G3) tend to progress to higher grades with recurrence. Grade 4 tumors (glioblastoma multiforme or GBM) commonly present as *primary* tumors (pGBM), with no prior history of occurrence at a lower grade. *Secondary* GBM (sGBM), on the other hand, has recurred in a patient previously diagnosed and treated for a lower grade [20]. Specific avenues of progression where astrocytoma manifests in G2 tumors that undergo transformation to the more aggressive G3 or GBM tumors, have been seen in both genetically engineered mouse models [21], as well as in humans [20]. My study included only GBMs with clear subtype

[1] This chapter includes material that was reproduced with permission from the following publication: C Wang, CC Funk, JA Eddy, ND Price (2013) Transcriptional Analysis of Aggressiveness and Heterogeneity across Grades of Astrocytomas. PLoS ONE 8(10): e76694. doi:10.1371/journal.pone.0076694 (all sections; text was collaboratively written with Cory Funk, James Eddy and Nathan Price).

designations (primary or secondary) and investigates differences between GBMs and lower grades as well as between these subtypes.

2.2 GENE EXPRESSION PROFILING: CHALLENGES AND STRATEGIES

A lot of transcriptomic studies have been done on human astrocytoma and differentially expressed genes and molecular signatures have been identified in previous microarray experiments, in an attempt to address clinical needs [22-27]. Unfortunately, as most of these studies were statistically underpowered, these signatures failed on independent validation sets, thus rendering them ineffectual [28]. Lab effect can obfuscate signal from noise in phenotypically similar tumors if sampled from different studies [29]. This can be overcome through use of multiple data sets when properly normalized—also minimizing the inherent biological noise [30]. Our present study adopted such a uniform approach to process raw expression data from multiple labs with one standard adjustment method, thereby increasing sample-to-sample correlation and decreasing heterogeneity across the data collected in different studies (**Figure 2.1A**).

Another strategy to mitigate biological noise is to analyze molecular profiles from individual genes or proteins in the context of biological network behaviors, and helps to link changes in gene expression to phenotype. Studying network behavior is especially relevant in cancer research as cancer stages and progression are marked by changes in network-level processes [31]. My research lab had previously developed a method called Differential Rank Conservation (DIRAC) [32], which measures the variation in *network ranking* (i.e., the relative ordering of genes from highest to lowest expression within a pre-defined network) among samples of the same phenotype and between samples of different phenotypes (**Figure 2.1C**). This enables evaluation of changes in gene expression at a network level based on relative expression between each of the network components, making the method independent of any normalization that does not affect rank (e.g., normalizing to total RNA, quantile normalization, etc.); additionally, the results do not depend on the other genes in the

transcriptome, meaning that it can be applied when only the genes in the network are measured.

After explaining the key challenges in transcriptomic study and my strategies to address the problem, I will present the major observations and improvements in the next section.

2.3 CONSENSUS PRE-PROCESSING REDUCES NOISE

Appropriate computational pre-processing is an important step in combined analyses of multi-site data to reduce technical variability between different studies. Consensus pre-processing, which normalizes raw expression data from multiple studies in a uniform manner, has been shown to reduce lab effects known to obfuscate biological signal when combining datasets from multiple labs [30]. Molecular signatures obtained after this step of processing have better prediction accuracy and lower variance than those from individual datasets. For example, average accuracy obtained training on four GBM datasets was considerably higher than training on individual GBM datasets [30]. We applied consensus pre-processing to the raw expression data for 336 patients collected from multiple independent studies. This greatly reduced sources of variation across studies, as measured by an increase in average sample-to-sample correlation from 81% to 91% (**Figure 2.2**). Reducing noise in the data enabled a more robust identification of variability across phenotypes.

2.4 GLOBAL DYSREGULATION OF NETWORKS

We first investigated global differences in network-level expression between astrocytoma grades by applying DIRAC to measure the rank conservation index of relative stability or consistency within each network ordering across a population [32]. If the orderings of genes within a specific network are mostly similar among different patients (i.e., highly conserved), the network is considered *consistent* within a phenotype. In the opposite case, more dissimilarity among patients is observed, and the network is considered *heterogeneous* or *dysregulated.* Extending this concept, averaging rank conservation indices over all networks provides a coarse measure of global regulation in different phenotypes.

We found that networks in normal brains are on average more highly conserved (0.957) than networks in advanced astrocytoma grades (G2, 0.937; G3, 0.930; and GBM (including both pGBM and sGBM), 0.915; $P < 0.001$ for ordering of phenotypes, based on one-way ANOVA) (**Figure 2.3A&2.3B**). In addition, global network rank conservation is significantly different between all pairs of phenotypes ($P < 0.05$, multiple pairwise t-tests). This trend demonstrates that more aggressive phenotypes have greater overall variation in network ordering among different samples. Increased genetic and cellular heterogeneity is a commonly recognized characteristic of highly malignant astrocytomas [33, 34]. GBM, the most malignant grade, is characterized by extensive heterogeneity as reflected in the moniker "multiforme," which derives from early histopathologic descriptions of a single tumor's highly varied morphologic features and connotes cellular heterogeneity [35]. Here, we show in a quantitative manner that *transcriptomic heterogeneity*, observed at the population scale, is generally correlated with increasingly aggressive phenotypes.

2.5 DIFFERENTIALLY REGULATED NETWORKS BETWEEN DISEASE STATES

Certain networks appear consistent in one phenotype but show drastically more sample-to-sample heterogeneity in another phenotype. Identifying the most differentially regulated networks can inform us about cellular processes and mechanisms most affected or perturbed from one disease state to another. We thus identified the most differentially regulated networks between normal samples and different astrocytoma grades as well as between different disease states (**Table 2.1**). For example, we identified 12 out of 248 networks that had a significant difference in conservation in comparing normal to G2 patients ($P < 0.01$ for each comparison, based on a binomial distribution; see **Chapter 2.9 Methods**); 10 out of these 12 networks showed increased dysregulation in G2. Similarly, in comparing G2 to G3, G3 to primary or secondary GBM, a strong majority of significantly dysregulated networks exhibited greater heterogeneity in the more malignant phenotype (**Figure 2.3**) ($\boldsymbol{P} < 0.01$ for each comparison, based on a binomial distribution). These quantitative results further support the idea that networks become increasingly dysregulated with increased malignancy.

2.5.1 DYSREGULATED NETWORKS IN G2 VS. NORMAL

Among the 12 significantly differentially regulated networks ($P < 0.01$) between G2 and normal brain, 5 networks (**PLCD**, **PLCE**, **AKAP13**, **CCR5**, and **ION)** are known to regulate protein kinase C (PKC) signaling and increase calcium release into the cell (**Figure 2.3**). Calcium signaling is a key player in neuronal transmission, microglia activation, and motility. Calcium signaling is especially crucial for transformed glioma cells to expand in the early stages of tumor development by sheer motility, as glioma cells cannot spread through the bloodstream [36]. Similarly, hyperactive PKC signaling is among the most distinguishing features of malignant brain tumors. PKC signaling stimulates both MAPK/ERK and PI3K/AKT pathways; it also supports degradation of extracellular matrices and allows for invasion of glioma cells [31].

Three networks, **ACETAMINOPHEN**, **SLRP**, and **PEPI**, mediating immune system responses, also showed increased dysregulation in G2 patients (**Figure 2.3C**). The **ACETAMINOPHEN** network was named after the commonly used drug Acetaminophen to reduce pain, targeting the cyclooxygenase enzymes. This network is also involved in inducing expansion of myeloid-derived suppressor cells (MDSC), which suppress T-cell responses to tumor growth [37]. Increased instability of this network may contribute to gliomagenesis by supporting development of MDSCs and their accumulation in the tumor microenvironment [38]. The **SLRP** network consists of 5 small leucine-rich proteoglycans (SLRPs), which are ligands of the Toll-like receptors responsible of regulating innate inflammatory response [39].

In contrast to the **ACETAMINOPHEN** and **SLRP** networks, **PEPI** showed significantly more consistent expression ordering in the cancer population (0.877 in normal and 0.945 in G2 patients, $P = 0.006$). This network activates neutrophils and generates the wound cleaning response—and is likely indicative of the normal physiological response to most tumors. In the early stage of forming malignant glioma cells, it is possible that some immune-related networks like **PEPI** act to prevent tumor cell migration and invasion through a more consistent expression program, while the dysregulation of other networks

like **SLRP** and **ACETAMINOPHEN** contributes to the immunosuppressive environment in the tumor.

2.5.2 Dysregulated networks in G3 vs. G2

Comparing network states in high-grade G3 to low-grade G2, all 5 networks with significant change in consistency of gene ordering showed greater heterogeneity in the more aggressive cancer grade (**Figure 2.3C**). **ERBB4** and **NOTCH** networks are part of the larger EGFR/ErbB signaling pathway. The key components in this pathway consist of four members of ErbB family of proteins (Erb1-4), which tend to form heterodimers and bind several cognate growth factors (e.g. EGF, TGF), activating downstream transcription factors (e.g. JUN, FOS) to regulate multiple cellular responses including proliferation and apoptosis [40]. This pathway has demonstrated substantial biological and transcriptional consequences such as activating downstream PI3K/AKT, PKC, and MAPK/ERK pathways. Up to 40% of GBMs display deletions in EGFR rendering it constitutively active, while others overexpress it through amplification or up-regulation of expression [41].

The **NOTCH** network interacts closely with EGFR to facilitate tumor angiogenesis. Our observation that **NOTCH** shows greater variability in expression ordering at the higher grade—from 0.908 (in G2 tumors) to 0.856 (in G3 tumors)—supports the hypothesis that it plays different roles in tumorigenesis of low-grade astrocytomas and high grade gliomas. That is, while inactive **NOTCH** functions as a tumor suppressor in low-grade G2 tumors, it is activated and may act as an oncogene in high grade astrocytomas, especially primary GBM [42].

The **TERT** network, responsible for telomerase activation, also showed greater dysregulation in G3 compared to G2. Telomerase activation and subsequent telomere maintenance are generally associated with the malignant transformation of normal cells to cancer cells [43]. The increased transcriptomic heterogeneity and network ordering inconsistency in higher grade astrocytomas further supports the known fact of telomerase dysregulation in malignant cancer phenotypes [44].

2.5.3 DYSREGULATED NETWORKS IN GBM VS. G3

We compared network conservation values between G3 and primary GBM and between G3 and secondary GBM separately, and obtained 38 and 16 differentially regulated networks, respectively. 13 networks appeared as significant in both comparisons ($P < 0.01$). GBM displays all the pathological features in the lower grades such as altered regulation in transcription and metabolism, calcium, and EGFR signaling (**Figure 2.3C**).

The **PLCD**, **PLC**, **TRKA**, and **HBX** networks all regulate release of intracellular calcium and function in similar ways as **PLCE** and **PKC**, identified in the lower grades. Notably, **HBX** includes 4 genes (GRB2, HRAS, SHC1, SOS1) that are part of the PI3K/AKT pathway—known to be hyperactivated in GBMs, resulting in uncontrolled cell growth, survival, proliferation, angiogenesis, and migration [45].

As expected, there are a number of networks involved in the complex EGFR regulatory pathway as in the other grades. The **CBL** network contains the ubiquitin ligase Cbl which degrades EGFR, thus down-regulates EGFR signaling [40, 46]; the **ERBB3** network likewise contains functionally similar components and plays a similar role in EGFR signaling. **TERC**, another network in this list, behaves like **TERT** to control telomerase regulation.

Interestingly, two networks (**LDL** and **S1P**) with critical roles in cholesterol metabolism also displayed significant dysregulation in GBM. **LDL** transports cholesterol, which is needed for cell membrane repair and synthesis, whereas **S1P** controls transcriptional regulation of cholesterol metabolism in response to cholesterol levels in the cell [47]. In addition, **S1P** connects to the earlier mentioned EGFR pathway through two sterol-regulatory element-binding proteins (SREBF1 and SREBF2) that are activated by PI3K. The interplay between S1P, SREB proteins and EGFR regulate the expression of fatty-acid synthase, which synthesize fatty acid and plays key role in cancer pathogenesis [48]. It has been reported the EGFR mutations (EGFRVIII) and PI3K promote tumor growth and survival through SREBP-1 dependent lipogenesis [49].

2.5.4 DYSREGULATED NETWORK DISTINGUISHING PGBM FROM SGBM

In comparing the two subtypes of GBM, primary to secondary GBM, it is interesting to note that the conservation value of **S1P** also decreased significantly from 0.812 in primary GBM to 0.769 in secondary GBM (P = 0.005). The SREBF1 gene in this network regulates and activates the IDH1 gene [50]. IDH mutations are commonly observed in lower grade and secondary GBMs but rarely in primary GBMs [51]. Thus, this network links IDH mutation to lipid homeostasis. Increased network dysregulation of **S1P** in secondary GBM offers quantitative support that IDH signaling is altered in this subset of GBMs.

2.6 MONOTONICALLY INCREASING AND DECREASING GENES IN ASTROCYTOMA PROGRESSION

Amidst the increased dysregulation of gene networks with increasing astrocytoma grade, we sought to identify instances where specific molecular changes—in this case, changes in expression of individual genes—occur in a unidirectional manner. We reasoned that such instances could provide insight into the oncogenic mechanisms or events that contribute to the pathology and/or transcriptomic heterogeneity found in astrocytoma. We therefore looked for genes whose expression level monotonically changed concomitant with increasing grade.

31 and 6 genes were found to decrease or increase their respective expression from normal to G2, G3, and GBM (**Figure 2.4, Table 2.2**). In evaluating DEGs between G3 and GBM, only genes differentially expressed in both pGBM *and* sGBM compared to G3 were included (see **Chapter 2.9 Methods**). We also tested for the statistical significance of the directionality of the genes (P < 0.001, see **Chapter 2.9 Methods**). The fact that specific genes change consistently with increasing astrocytoma grade may reflect shared oncogenic mechanisms among phenotypically similar tumors. Interestingly, similar to the differentially regulated networks, several of these genes identified are also associated with key processes such as calcium signaling and metabolism and/or are located in the endoplasmic reticulum (ER) or mitochondria. The commonalities shared by gene-based and network-based analysis

may represent potential connections between genetic heterogeneity at the tumor level and expression heterogeneity at the population level.

The significance of calcium signaling and metabolic genes may relate to how cells respond to additional metabolic requirements needed for tumor cell division and cell cycle progression with increased aggressiveness. At the same time, cells that ultimately constitute the tumor mass have been selected for their ability to avoid apoptosis while facilitating the increased metabolic flux. As such, we see genes implicated in regulation of apoptosis. We discuss in detail below how representative genes are involved in the above-mentioned functional categories and how they interact to bring about changes reflective of astrocytoma pathology. A summary of the genes and their respective functions is shown in **Figure 2.5**.

2.6.1 GENES IMPLICATED IN CALCIUM SIGNALING AND/OR APOPTOSIS

Among the monotonically changing genes, TMEM66, STRN3, CANX, and CPEB3 are known to affect calcium and apoptotic signaling; all of them, with the exception of CANX, showed decreased expression with increasing tumor grade.

TMEM66, also known as SARAF, is localized to the ER lumen and affects calcium storage [52]. Following calcium release from the ER, calcium stores are replenished through calcium release activated channels (CRAC) to re-enter the ER lumen [53]. Decreased SARAF, as we observed in our study, would potentially lead to an inability to close the CRAC channels and disrupt calcium homeostasis in aggressive gliomas.

Striatin, calmodulin binding protein 3 (STRN3) is another monotonically decreasing gene and participates in apoptosis and calcium release. It is found to be both cytosolic and membrane-bound and is expressed primarily in the brain and muscle [54]. STRN3 binds with calmodulin in the presence of calcium [55]. It reacts with protein phosphatase 2a (PP2a), which, along with the promyelocytic leukemia (PML) protein, stimulates IP_3R-mediated Ca^{2+} release from ER. PML modulates calcium-mediated apoptotic stimuli through binding with PP2a and IP_3R [56]. Decreased expression of STRN3 in aggressive gliomas likely reflects changed apoptotic calcium signaling mechanisms in these tumors.

Cytoplasmic polyadenylation element binding protein 3 (CPEB3) is a nucleocytoplasmic shuttling RNA-binding protein. It is involved in both calcium signaling and EGFR degradation. CPEB3 inhibits EGFR expression by preventing the translation of STAT5B, a regulator of EGFR transcription [57]. As a monotonically decreasing gene, lower expression of CPEB3 would similarly lead to an increase in EGFR. Notably, CPEB3 is located on chr10q 23.32, very close to the locus of PTEN (chr10q 23.31). Loss of this region is known to occur in several cancers [58] and it is conceivable that loss of CPEB3 contributes to altered EGFR signaling along with PTEN loss.

In contrast to the above three genes, calnexin (CANX) was found to increase at the mRNA level with increased astrocytoma grade. CANX is an ER chaperone protein that binds with free calcium ions. It is a critical component of the mitochondria associated membrane (MAM), with over 80% of it located in the MAM, along with the aforementioned STRN3-associated protein PP2A. CANX regulates the activity of sarcoplasmic/endoplasmic reticulum calcium ATPase (SERCA) by acting as a calcium buffer in the MAM [59]. Depending on its palmitoylation status, CANX shuttles between the ER and MAM [60].

2.6.2 GENES IMPLICATED IN METABOLISM AND MITOCHONDRIA

A few monotonically changing genes identified have metabolic functions. Proteins encoded by these genes sit closely to each other in the mitochondria, which is responsible for essential cellular processes such as energy production, storage of calcium ions, and cell death. In recent years, there has been increased reports of the role of mitochondria in calcium signaling [61], which helps to connect mitochondrial metabolic genes to calcium signaling. Metabolic regulation of calcium in mitochondria is mediated through the effects of dehydrogenases. Calcium ions activate matrix dehydrogenases, increase available NADH and electrons for the respiratory chain, and eventually accelerate ATP production [62].

NDUFB8 and NDUFB1 are two monotonically decreasing genes that encode subunits of respiratory chain NADH dehydrogenase complex I. Decreased expression of these proteins causes respiratory chain dysfunction, reduces the driving force for calcium transfer and

available electrons in the respiratory chain, decreasing ATP production. This observation may reflect a reduction of mitochondrial ATP synthesis via oxidative phosphorylation—contributing to the Warburg Effect; as the tumor grows to more aggressive stages, the metabolism of proliferating tumor cells is adapted to proliferation mechanism rather than efficient ATP production [63]. Interestingly, the NDUFB8 gene is also found very close to PTEN (locus of NDUF8: chr10q 24.31 and PTEN, chr10q 23.31).

ACSL4 (acyl-CoA synthetase), a monotonically decreasing gene, converts fatty acids to fatty esters and plays an important role in lipid metabolism. Similar to CANX, ACSL4 is also found in MAM, which is a critical metabolic hub in lipid metabolism [64]. Though normally recognized as a metabolic gene, ACSL4 regulates synaptic vesicles along axons. Knockout of ACSL4 in embryonic stem cells was shown to significantly reduce neuronal differentiation [65]. A de-differentiated neuronal state in higher-grade tumors resembles how neural stem cells display higher potential of proliferation and angiogenesis [66].

Other mitochondrial genes involved in metabolism include ornithine aminotransferase (OAT), a monotonically decreasing gene that converts arginine and ornithine into neurotransmitters glutamate and GABA. GABA receptors and glutamate transporters have been reported to be down-regulated in brain tumors [67]. Another mitochondrial gene identified is n-myristoyltransferase 2 (NMT2) which plays a role in protein myristoylation, proliferation, and apoptosis [68].

2.7 DIRAC-BASED CLASSIFICATION IDENTIFIES ACCURATE NETWORK SIGNATURES FOR DISTINGUISHING GRADES

The high degree of transcriptomic heterogeneity observed in increasingly aggressive astrocytoma tumors creates substantial variance when searching for robust molecular signatures between grades. Still, identifying such signatures is critical to elucidating mechanistic differences between more and less aggressive tumors. Network-based approaches such as DIRAC are advantageous for extracting signal from noise, as the patterns of functional groups might be less *within the same phenotype* than those of

individual genes. Furthermore, DIRAC quantifies the relationships between genes, and operates on these pair-wise expression patterns within networks, thereby reducing the impact of noisy changes in single gene expression. Using DIRAC, we compared each of the four phenotypes to all other phenotypes (e.g., normal brain against G2, G3, GBM; G2 against G3 and GBM; etc.) (**Figure 2.6**).

We were able to clearly separate normal brain tissues from G2, G3, and primary GBM. Furthermore, these tumors could be distinguished from each other with good accuracies (> 80%, except 78% in the case of G3 and pGBM). In separating secondary GBM from other grades, however, classification signals are not as strong (average accuracy of primary GBM and secondary GBM vs. all other phenotypes are 86% and 77%, respectively). This difference in classification performance very likely reflects the fact that secondary GBMs are derived from lower grades and therefore share more common genomic and transcriptomic characteristics in their expression profiles compared to primary GBMs, which develop spontaneously and display more pronounced phenotypic differences from other grades.

It was also difficult to separate primary and secondary GBMs (accuracy 69%) based on their transcriptomes, even though they are known to develop from separate genetic pathways [69]. They are indistinguishable by histology, as both share the same histological grade [70]. Both subtypes share a number of genomic and transcriptomic similarities such as LOH on chromosome 10q and deregulation of the PI3K/ATK pathway [69]. Another reason for the relatively lower accuracy is possibly due to the signal present from other subtypes such as proneural (PN), mesenchymal, or proliferative subtypes (the latter two collectively known as non-PN) within GBMs, which appear to be more distinct than the transcriptomic differences we observe between primary and secondary GBM. In terms of survival, the PN subtype is reported to be less aggressive than other subtypes [25]. In support of this hypothesis, we applied DIRAC on a subset of GBM with known PN/non-PN designations, and separated the proneural subtype from the rest with an accuracy of 78%. This accuracy being higher than for the separation of pGBM and sGBM suggests that molecular subclasses in glioblastomas may look more different than traditional pGBM/sGBM classes, especially

in the context of network behavior; hence DIRAC detected the stronger classification signals more easily. The best 10 network-based classifiers selected by DIRAC to separate tumor samples from normal brains are listed in **Table 2.3**. In each pair-wise comparison, we included different metrics (sensitivity, specificity, and accuracy) and group size information to demonstrate the ability of DIRAC to distinguish different grades of brain tumors (**Table 2.4**).

2.8 CONCLUSIONS

We report here a systems approach to investigate molecular changes underlying astrocytoma pathology. Leveraging a large cohort of publicly available gene expression data sets, we have conducted the first meta-analysis that examines together the transcriptomes of three astrocytoma grades along with corresponding normal samples. We have combined individual gene- and network-based approaches to identify meaningful patterns of expression within and between different grades. The trend we observed of greater network dysregulation with higher grade represents a quantified measure of increasing inter-patient transcriptomic heterogeneity in more aggressive astrocytomas. We also identified genes that exhibit monotonically increasing or decreasing expression with increased grade—these genes are potentially reflective of shared oncogenic mechanisms among phenotypically similar tumors. Notably, monotonically increasing or decreasing changes in gene expression, parallel to increasing network dysregulation, presents a putative bridge between the known genetic heterogeneity of astrocytomas and expression heterogeneity at the population level, as analyzed in this meta-study.

Additionally, we identified networks distinguishing different astrocytoma grades from normal as well as network markers separating between glioma grades. This work presents significant results that enable better characterization of different human astrocytoma grades, and hopefully will lead to improvements in diagnosis and therapy choices.

2.9 METHODS

2.9.1 COLLECTION AND INTEGRATION OF TRANSCRIPTOMIC DATA

Raw microarray CEL files from previous studies were compiled from the NCBI Gene Expression Omnibus (GEO). We used data collected from the most abundant source platform currently, Affymetrix HG-U133A or its complimentary version, HG-U133-Plus 2.0 GeneChips (Affymetrix, Santa Clara, CA). **Table 2.5** lists the GEO accession number, year of publication, and the number and grades of samples reported in each original study.

A "consensus pre-processing" method was applied to the CEL files to normalize differences introduced by non-uniform studies and sample preparation procedures. This method is described in greater detail in [30] and was used in that study to demonstrate that classifiers performed better on novel datasets when trained on multiple, integrated, pre-processed datasets. Briefly, common probe sets (22,277) shared by the two platforms (U133A and U133-Plus 2.0) were identified according Affymetrix descriptions, and GeneChip RMA (GC-RMA) normalization was applied to raw expression data for these probes across all microarray samples [71]. GC-RMA was implemented in the Matlab Bioinformatics Toolbox with the threshold for presence defined based on prior studies from Affymetrix [72]. Probes having 0% present calls for any phenotype were removed. Following these criteria, 15,827 probes were kept for further analysis.

When converting the probe intensity matrix to a gene expression matrix, probes that mapped to multiple genes were eliminated to remove ambiguity. For multiple probes corresponding to the same gene, the maximum intensity was used. Finally, all absolute intensity values were replaced by their relative ranks within each array.

2.9.2 COMPUTATION OF RANK CONSERVATION INDICES IN DIRAC

For all network analyses performed with DIRAC, expression levels of genes were grouped into 248 human signaling networks, defined according to the BioCarta gene sets collection

in the Molecular Signatures Database (MSigDB) [73]. For each selected network, we used DIRAC to compute the expected ordering of network genes (rank template) for each phenotype, and we subsequently measured how closely each sample's network ordering matched the phenotype-specific template (rank matching score). The rank conservation index, calculated by averaging rank matching scores across samples in a phenotype, indicates how consistently each network is ordered within a population. Averaging the rank conservation indices over all networks for a phenotype provided a single value estimating the relative heterogeneity or *dysregulation* of networks for that phenotype.

2.9.3 IDENTIFICATION OF MOST DIFFERENTIALLY REGULATED NETWORKS ACROSS GRADES

The difference in rank conservation indices between two phenotypes (e.g., normal vs. cancer or lower grade vs. higher grade) was calculated for each network. Networks were ranked based on the magnitude of the difference. To establish statistical significance, the original phenotype labels were permuted and randomly assigned to samples, and the absolute difference in rank conservation indices was calculated for all networks in each phenotype. These steps were repeated for 1,000 permutations to generate a null distribution of rank conservation differences, and the significance level for each difference was measured as the probability of observing the same fraction or higher at random.

2.9.4 IDENTIFICATION OF MONOTONICALLY CHANGING GENES

We selected differentially expressed genes (DEGs) for each adjacent pair of astrocytoma grades, (control vs. G2, G2 vs. G3, etc.) based on the Wilcoxon rank-sum test ($P < 0.05$ after Bonferroni correction). In the intersection of these DEG sets, genes with monotonically increasing ranks were defined as *increasing* genes, and monotonically decreasing ranks as *decreasing* genes.

In order to test the robustness of these monotonically changing genes, we randomly selected 80% of all samples in each phenotype, and with this subset of samples, we tracked whether genes were similarly increasing or decreasing across grades as they were with the full set of samples. We repeated this selection process 1000 times and recorded how often the

identified genes appear with the same pattern. Genes that appeared at least 500 times in 1000 permutation tests were considered as high confidence genes and used for subsequent analysis.

In order to test the significance of the directionality of genes, the original phenotype labels were permuted and randomly assigned to samples, and the number of monotonically changing genes in each permutation was calculated for both the increasing and decreasing case. The procedure was repeated for 1000 times and a null distribution for the gradation of genes was established. A *P*-value for the directionality/trend of the genes was assigned based on the probability of observing the same number of genes at random.

2.9.5 CLASSIFICATION OF DISEASE PHENOTYPES WITH DIRAC

In addition to conservation of network ordering within a phenotype (measured by the rank conservation index), DIRAC can also be used to identify networks ordered differently (variably expressed) between two phenotypes. Rank matching scores were calculated for each class, and predicted class labels were assigned based on similarity of each patient's individual profile to either of the two templates. Apparent accuracy for classification with these predicted class labels was then calculated for all networks [32]. A null distribution of network classification rates was generated by randomly permuting phenotype labels 1000 times, and the significance level was measured as the probability of observing classification rates. To address the issue of multiple-hypothesis testing, the corresponding false discovery rate (FDR) was calculated for each significance level, representing the fraction of expected false positives at any defined cutoff [32]. We used leave-one-out cross validation to estimate the error rate of DIRAC-based classification for each pair of phenotypes.

2.10 CHAPTER 2 FIGURES AND TABLES

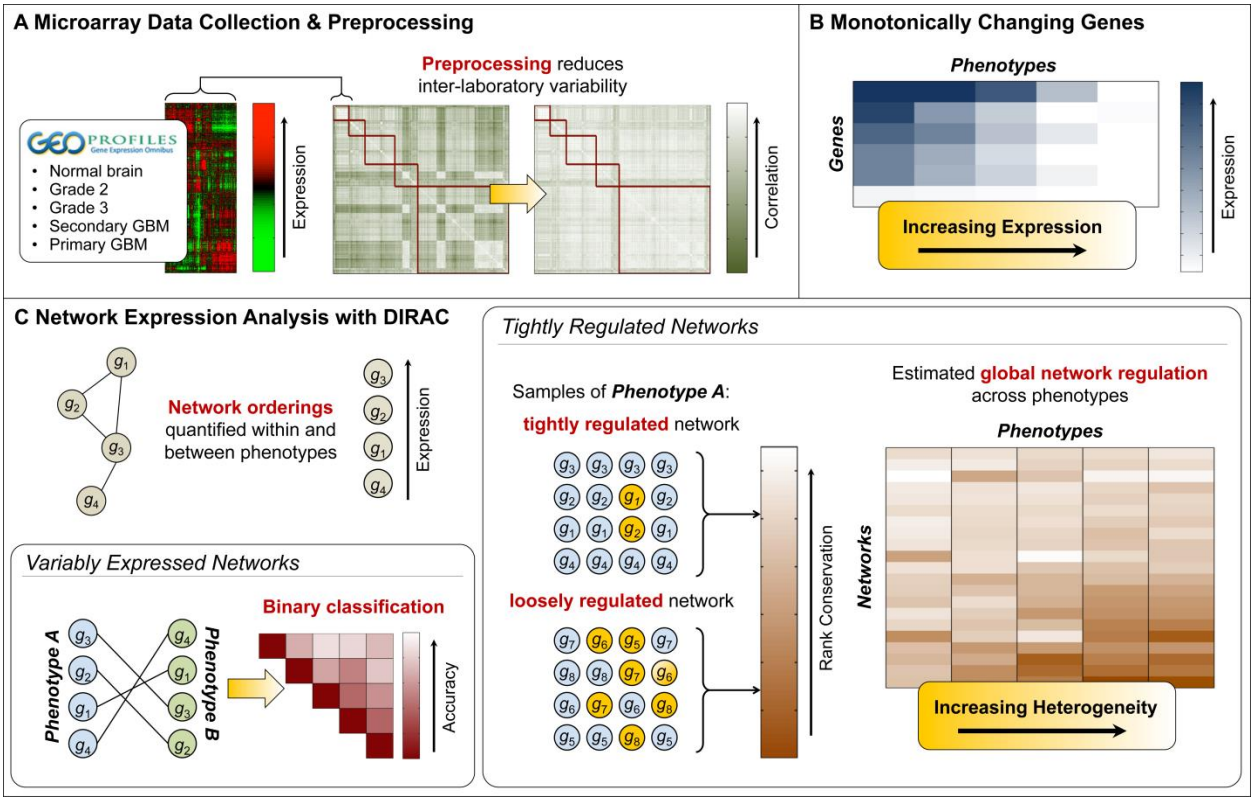


Figure 2.1 Overview of approach.

A) We minimized experimental variation due to lab effects by performing uniform pre-processing. **B)** Genes that either monotonically increased or decreased in parallel with increasing astrocytoma grade were identified. **C)** Molecular signatures that can accurately distinguish between different grades were established using Differential Rank Conservation (DIRAC). We also examined broad patterns of network regulation across all astrocytoma grades.

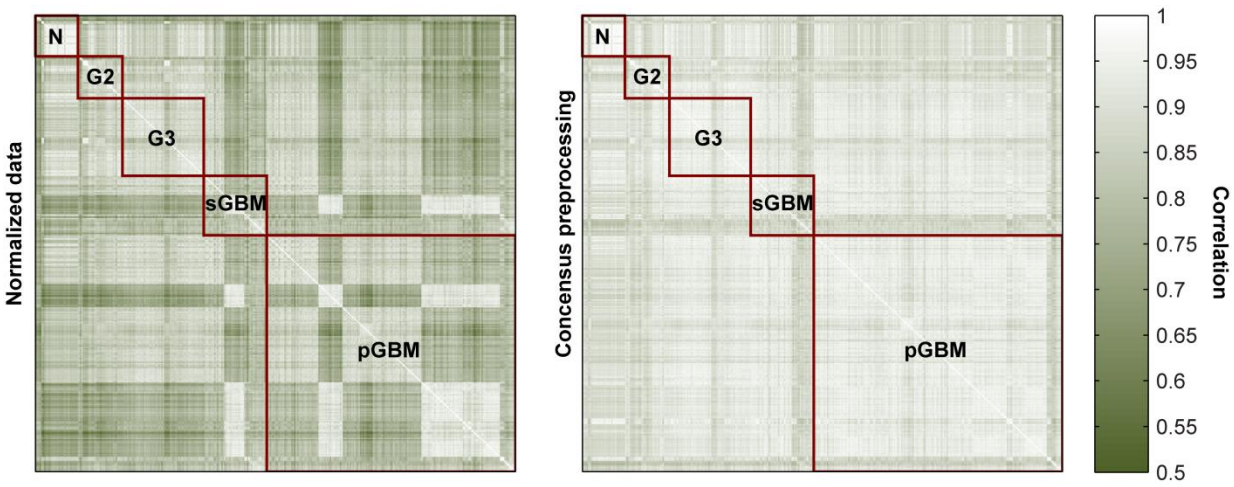


Figure 2.2 Pearson-correlation matrix before and after consensus pre-processing

The heatmaps display correlation coefficients among all samples included in this study. The axes represent sample numbers. In the left figure, the purple borderlines of each box delineate different phenotypes, which coincide with the sample batches. Samples from the same laboratories or studies showed higher homogeneity than other samples. On the other hand, in the right figure, laboratory effects are much less obvious; tumor samples across different studies or phenotypes all look highly correlated with average correlation coefficient increased from 0.81 to 0.91.

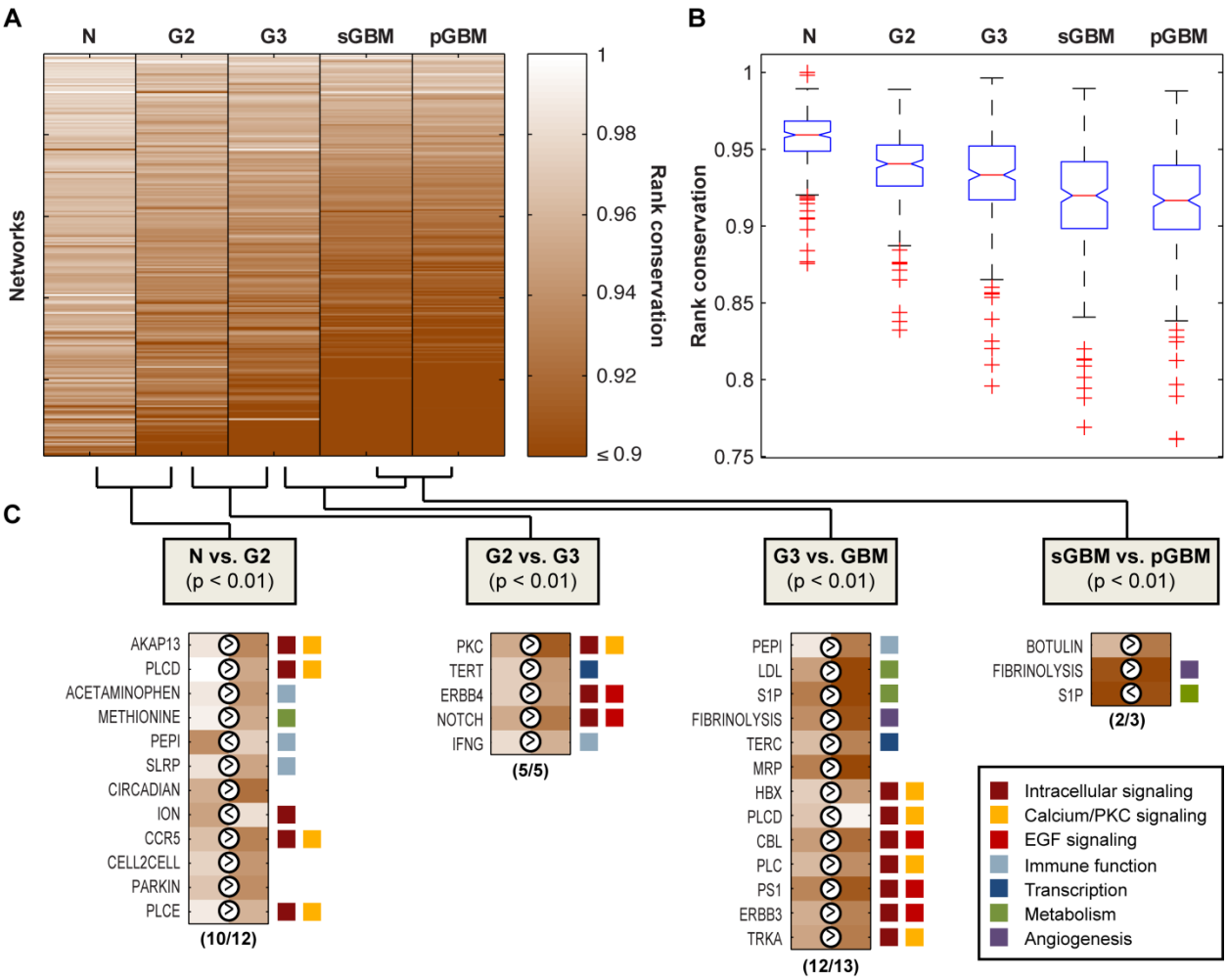


Figure 2.3 Network-level expression heterogeneity across tumor grades

A) Global trend of network regulation level decreases with increeasing grade. The vertical axis represents examined networks, while the horizontal axis represents five phenotypes. Colors represent rank conservation indices for each network. Light colors indicate high consistency of network ranking in a phenotype and the dark colors indicate large heterogeneity of networks. Networks in sGBM and pGBM tumors become much more heterogeneous compared to the normal cases. **B)** One-way ANOVA comparing the mean rank conservation values of different phenotypes. **C)** A list of most deregulated networks between adjacent tumor grades and their major biological functions. The ">" and "<" indicate the magnitude of network regulation. For instance, AKAP13 has a larger rank conservation index in normal samples and thus is more regulated in normal compared to G2.

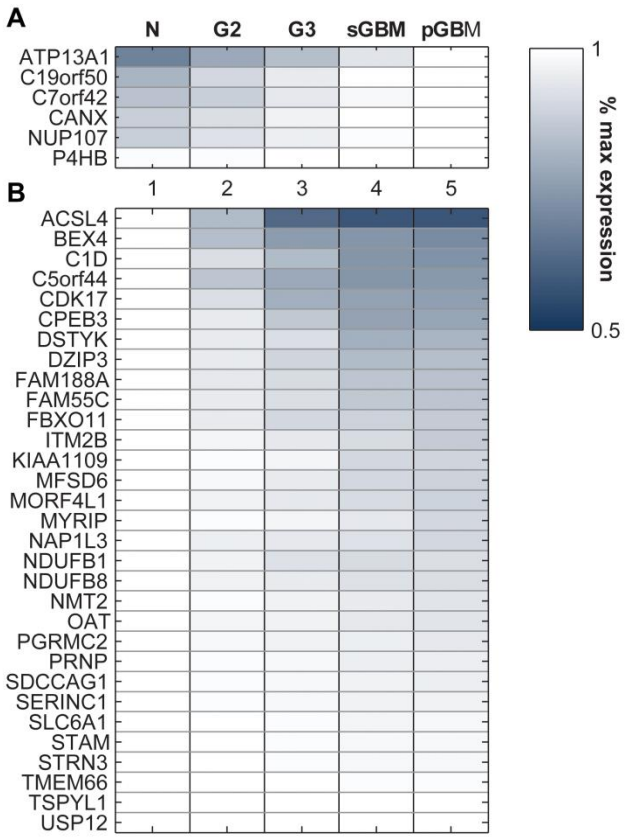


Figure 2.4 Genes showing consistent dysregulation with progression

Colors on the heatmap represent relative expression values of genes in different phenotypes. The vertical axis lists the differentially expressed genes and the horizontal axis lists the phenotypes. All expression values are normalized as the percentage of maximum expression value for the gene across all phenotypes. For the up-regulated genes in panel **A)** the maximum expression is either sGBM or pGBM, so we see all genes have brightest color in these two phenotypes; similarly the down-regulated genes in panel **B)** decrease their expression systematically from normal to GBM, and the intensity level also increases with grade.

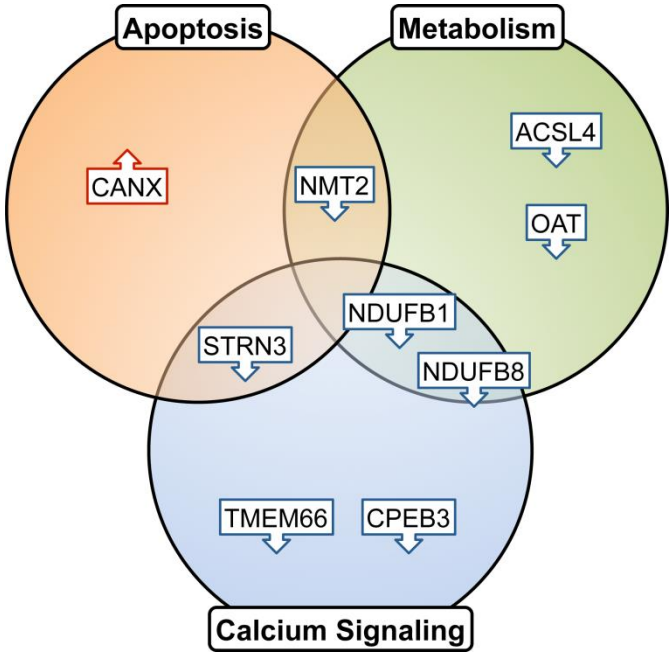


Figure 2.5 Functional categories among monotonically changing genes

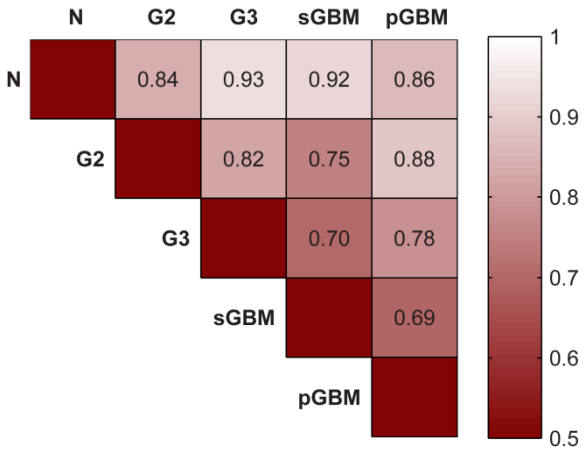


Figure 2.6 Classification accuracy with biocarta database networks

This heatmap displays leave-one-out cross-validation accuracies of DIRAC-based classifications on each phenotype vs. all other phenotypes. DIRAC could distinguish more distant grades like normal vs. GBMs; it is especially hard to separate G3 or pGBM from sGBM.

Biocarta Network	Rank Conservation in normal	Rank conservation in G2
AKAP13	0.97	0.871
PLCD	1	0.905
ACETAMINOPHEN	0.975	0.89
METHIONINE	0.983	0.909
PEPI	0.877	0.945
SLRP	0.968	0.9
ION	0.898	0.964
CIRCADIAN	0.91	0.844
CCR5	0.93	0.865
CELL2CELL	0.951	0.887
PLCE	0.976	0.916
PARKIN	0.936	0.876

Biocarta Network	Rank conservation in G2	Rank conservation in G3
PKC	0.909	0.825
TERT	0.945	0.889
ERBB4	0.946	0.891
NOTCH	0.908	0.856
IFNG	0.962	0.914

Table 2.1A&B: A) (left) Most differentially deregulated networks between normal and G2 patients B) (right) Most differentially deregulated networks between G2 and G3 patients

Biocarta Network	Rank Conservation in G3	Rank conservation in GBM
CBL	0.891	0.841
ERBB3	0.911	0.865
FIBRINOLYSIS	0.876	0.813
HBX	0.946	0.893
LDL	0.896	0.794
MRP	0.86	0.801
PEPI	0.973	0.858
PLCD	0.936	0.987
PLC	0.924	0.877
PS1	0.867	0.82
S1P	0.857	0.769
TERC	0.926	0.865
TRKA	0.903	0.857

Biocarta Network	Rank Conservation in primary GBM	Rank conservation in secondary GBM
BOTULIN	0.856	0.921
FIBRINOLYSIS	0.762	0.813
S1P	0.812	0.769

Table 2.1C&D: C) (left) Most differentially deregulated networks between normal and G2 patients D) (right) Most differentially deregulated networks between G2 and G3 patients

Gene Name	Chromosome Locus	Putative gene functions
ATP13A1	19p13.11	ATP binding
C19orf50	19p13.11	Vesicle transport
C7orf42	7q11.21	Unknown function
CANX	5q35	Calcium signaling, associates with MAM
NUP107	12q15	Inhibits apoptosis
P4HB	17q25	Hydroxylation of prolyl residues in preprocollagen

Table 2.2A monotonically increasing genes and their putative functions

Gene Name	Chromosome Locus	Putative gene functions
ACSL4	Xq22.3-q23	Fatty acid metabolism
BEX4	Xq22.1-q22.3	Brain expressed X-linked gene
C1D	2p13-p12	Calcium signaling
C5orf44	5q12.3	Possible member of the TRAPP complex
CDK17	12q23.1	Ubiquitin
CPEB3	10q23.32	Activated by Calcium, degradated by EGFR
DSTYK	1q32.1	Extrinsic apoptotic signaling
DZIP3	3q13.13	E3 Ubiquitin ligase
FAM188A	10p13	Ubiqutin plays a role in apoptosis
FAM55C	3q12.3	Belongs to the neurexophilin and PC-esterase domain family; Unknown function
FBXO11	2p16.3	Subunit of ubiquitin protein ligase complex
ITM2B	13q14.3	Encodes a transmembrane protein which inhibits the deposition of beta-amyloid.
KIAA1109	4q27	Associated with spermatocyte and adipocyte differentiation
MFSD6	2q32.2	Transmembrane transport
MORF4L1	15q24	Transcriptional regulation
MYRIP	3p22.1	Vesicle transport, functions as a AKAP
NAP1L3	Xq21.3-q22	Linked closely a region of genes responsible for several X-linked mental retardation syndromes
NDUFB1	14q32.12	Mitochondrial metabolism
NDUFB8	10q24.31	Mitochondrial metabolism
NMT2	10p13	Catalyzes the reaction of N-terminal myristoylation of many signaling proteins
OAT	10q26	Mitochondrial protein, involves in

Table 2.2B

		the production of proline from ornithine
PGRMC2	4q26	Modulate cytochrome P450 enzymes
PRNP	20p13	Involved in neuronal development and synaptic plasticity, implicated in neurodegenerative diseases
SDCCAG1	14q22	p53-dependent and -independent DNA damage-induced apoptosis
SERINC1	6q22.31	Mitochondrial metabolism
SLC6A1	3p25.3	GABA transporters
STAM	10p14-p13	Vesicle transport, functions as a AKAP
STRN3	14q13-q21	Functions as scaffolding or signaling protein
TMEM66	8p12	Calcium signaling
TSPYL1	6q22.1	Implicated in X-linked pathologies
USP12	13q12.13	Targets Notch receptor for deubiquitination

Table 2.2B (cont.) monotonically decreasing genes and their putative functions

BioCarta Network	Apparent Accuracy	BioCarta Network	Apparent Accuracy
PTDINS	0.918	MCALPAIN	0.966
EGF	0.918	PTDINS	0.945
FAS	0.918	ERK	0.940
TNFR1	0.918	G2	0.932
RACCYCD	0.902	G1	0.932
CBL	0.902	EGF	0.932
PDGF	0.902	CELLCYCLE	0.925
IL1R	0.902	PROTEASOME	0.924
ACTINY	0.902	RACCYCD	0.924
HIVNEF	0.902	AT1R	0.916
BioCarta Network	**Apparent Accuracy**	**BioCarta Network**	**Apparent Accuracy**
CELL2CELL	0.932	HIVNEF	0.955
P38MAPK	0.921	STRESS	0.944
G2	0.913	CHEMICAL	0.940
G1	0.911	DEATH	0.937
ERK	0.910	MCALPAIN	0.929
MPR	0.910	P38MAPK	0.926
PROTEASOME	0.908	G2	0.924
VEGF	0.905	IL2RB	0.924
CELLCYCLE	0.905	MPR	0.924
HIVNEF	0.904	CELLCYCLE	0.921

Table 2.3 Top networks selected by DIRAC to classify tumor grades vs. normal brains (*P-value* < 0.0001). Top 10 networks for G2 vs. N (top left), G3 vs. N (top right), sGBM vs. N (bottom left), and pGBM vs. N (bottom right).

Class 1	size	percentage	Class 2	size	percentage	sensitivity	specificity	accuracy
N	30	0.492	G2	31	0.508	0.833	0.833	0.833
N	30	0.345	G3	57	0.655	0.947	0.917	0.932
N	30	0.147	pGBM	174	0.853	0.937	0.9	0.918
N	30	0.405	sGBM	44	0.595	0.841	0.883	0.862
G2	31	0.352	G3	57	0.648	0.807	0.839	0.823
G2	31	0.151	pGBM	174	0.849	0.891	0.871	0.881
G2	31	0.413	sGBM	44	0.587	0.705	0.79	0.747
G3	57	0.247	pGBM	174	0.753	0.784	0.781	0.783
G3	57	0.564	sGBM	44	0.436	0.716	0.684	0.7
pGBM	174	0.798	sGBM	44	0.202	0.718	0.659	0.689

Table 2.4 Sensitivity, specificity and accuracy of each classification

Formula used to calculate each metric

Sensitivity = TP / (TP + FN)

Specificity = TN / (TN + FP)

Accuracy = 0.5* TP/ (TP + FP) + 0.5* TN/ (FN +TN)

TP: true positive; FN: false negative; TN: true negative; FP: false positive

Platform	Authors of Study (year, GSE accession)	Number of patients in each class				
		Normal	Grade 2	Grade 3	Primary GBM	Secondary GBM
U133A	Freije *et al* (2006, GSE 4412)[24]	0	0	8	34	12
	Phillips *et al* (2006, GSE 4271) [25]	0	0	21	55	0
	Wong *et al* (2008, GSE 12907) [26]	4[a]	0	0	0	0
	Rich *et al* (2005, GSE 13041) [74]	0	0	0	31	0
	Lee *et al* (2008, GSE 13041) [75]	0	0	0	15	13
	Barrow *et al* (2008, GSE 13041) [75]	0	0	0	28	3
	Mcdonald *et al* (2005, GSE 3185)	0	3	0	0	0
Total U133A		4	3	29	163	28
U133-Plus 2.0	Sun *et al* (2006, GSE 4290) [27]	23[b]	7	19	0	0
	Liu *et al* (2010, GSE 19728)	1[c]	5	5	0	0
	Lee *et al* (2008, GSE 13041)[75]	0	0	0	11	16
	Turkheimer et al (2006,GSE 2817) [76]	0	6	0	0	0
	Chow et al (2010, GSE 22927) [77]	0	6	0	0	0
	Grzmil et al (2011, GSE 15824) [78]	2	4	4	0	0
Total U133-Plus 2.0		26	28	28	11	44
Total		***30***	***31***	***57***	***174***	***44***

Table 2.5. Summary of microarray expression datasets included in the study

[a] Includes one normal fetal brain RNA, one normal cerebellum RNA, and two normal tissues surgically removed tissue adjacent to resected tumor tissue and RNA extracted

[b] Brain samples of epilepsy patients

[c] Pooled normal brain tissue

CHAPTER 3 IDENTIFICATION OF PROGNOSTIC MARKERS FOR HIGH-GRADE ASTROCYTOMAS

3.1 HETEROGENEITY AND PROGNOSIS IN HIGH-GRADE ASTROCYTOMA

Heterogeneity in high-grade astrocytomas (HGAs; i.e., G3 and GBM) also manifests as differences in clinical outcome, which are often difficult to predict. Because of this, much effort has been devoted to identifying *subtypes* within a large collection of samples. A long-standing approach to classify HGAs is based on clinical histories of patients: primary tumors tend to occur in older patients, and correspond with slightly shorter survival time than secondary tumors [20]. Another approach by Phillips *et al.*, utilized a set of 35 genes to group HGAs into three subclasses, each resembling a corresponding stage in neurogenesis. One subclass (proneural or *PN*), exhibiting longer average survival, contained neuronal lineage markers; the two other tumor classes (proliferative and mesenchymal, collectively known as non-proneural or *non-PN*) were enriched for neural stem cell markers and had short survival times [25].

This chapter presents my work on using an extension of DIRAC-based classification to identify networks capable of predicting subtypes with differential survival in HGA. We found that variable expression of an erythropoietin network, which is known to mediate neuroprotection through NF-κB signaling (EPONFκB), could efficiently separate HGA patients into two groups with a significant survival difference. The prognostic value of this network was enhanced when combined with an established scheme for separating HGA patients (i.e., proneural/non-proneural) [25]. Two classes separated by both markers differed more significantly than classes defined by histologically determined grades or by proneural/non-proneural status alone.

3.2 THE EPONFκB NETWORK EXHIBITS PROGNOSTIC VALUE

To further explore the effects of heterogeneity among more aggressive astrocytoma tumors, we aimed to identify subpopulations within combined G3 and GBM samples (HGAs) with different clinical outcomes. Perhaps the most clinically relevant metric to evaluate subtypes is patient survival; this has been previously explored for primary vs. secondary tumors and the proneural (*PN*) vs. non-proneural (non-*PN*) subtypes (described above). With available survival information from 239 patients, I found that the EPONFκB network showed significant prognostic value. Using a distance matrix based on the DIRAC metric (see **Chapter 3.3 Methods**), the genomic profiles of all HGA patients were grouped into two clusters using unsupervised clustering. Subsequent log-rank tests on the survival estimates of these two groups (*EPOLONG* and *EPOSHORT*, indicate longer or shorter survival times) gave a *P*-value of $1.8e^{-5}$ (corrected for multiple hypothesis testing) (**Figure 3.1**), outperforming separation by path of progression (primary vs. secondary tumors, $P = 0.002$). Besides this network, four other BioCarta-defined networks showed significant *P*-values in their respective log-rank tests (**Table 3.1**).

I next sought to improve prognostic predictions by combining the EPONFκB marker with established subtyping schemes. Using the previously reported proneural signature on microarray samples collected and processed from GEO, *PN* patients can be seen to survive longer than non-*PN* patients ($P = 5.7e^{-8}$) (**Figure 3.1**). However, the 73 patients labeled both as *PN* and *EPOLONG* differed from the 68 patients labeled both as *non-PN* and *EPOSHORT* with the highest significance ($2.4e^{-10}$). This integrated method of clustering HGA patients also outperformed histological separation (G3 vs. GBM tumors, $P = 7.7e^{-6}$) (**Figure 3.1**).

Nearly all G3 specimens (20/21 or 95%) were labeled as both *PN* and *EPOLONG*, indicating both markers having similar powers in distinguishing histologically less aggressive populations. The *EPOLONG* subclass comprises significant numbers of both *PN* (73/161 or 45%) and *non-PN* (88/161 or 55%) subclasses, indicating that these two markers

identify different populations of patients in HGA. It is interesting to note that *EPOLONG* still contains a higher proportion of *PN* (45%) patients than *EPOSHORT* (10/78 or 13%).

Involvement of the EPONFκB network in gliomas is an interesting and highly debated subject because of its relevance to clinical treatment. Standard treatment of GBM with radiation does significant harm to the surrounding brain, resulting in significant collateral damage. This damage is referred to as "radiochemobrain" and results in slowing psychomotor skills, cognitive decline, fatigue, and loss of drive, all of which significantly reduce the quality of life [74]. To counteract these effects, patients are sometimes given hematopoetic growth factor erythropoietin (EPO) prior to and following radiation. EPO signaling cross-activates the anti-apoptotic transcription factor NF-κB, and mediates neuroprotection against oxidative stress. EPO has pleiotrophic effects on the brain including anti-apoptotic, antioxidative, neurotrophic, axon-protective, angiogenic, and neurogenic—many of which are associated with neuroprotection against the side effects of radiation and chemotherapy [75-78]. In addition, EPO has also been shown to improve the responsiveness of tumors to radiation therapy in human glioma xenographs by increasing tumor oxygenation [79].

For all the positive effects EPO is believed to have, some have argued that these same effects could potentially promote tumor growth. Recently, EPO signaling was shown to be involved in angiogenesis of human glioma cells as well as cancer stem cell maintenance [80]. Still, others have shown that while EPO does augment the survival of glioma cells, it is unlikely to appreciably influence basal glioma growth [81]. While my results implicate EPONFκB as a novel predictor of patient survival, it is unclear if the observed modulation reflects increased, decreased or variable network signaling.

3.3 CONCLUSIONS

I discovered a signature predictive of survival in HGA patients based on the gene rankings in the EPONFκB network. Strikingly, combining the EPONFκB network and previously reported signatures outperformed histology-based grading or those separated solely based on

proneural/non-proneural status as predictors of survival. The identification of the EPONFκB network as a potential prognostic factor demonstrates the utility of deriving molecular diagnostic signatures from multiple studies. Ultimately, the results of this analysis could lead to improvements in diagnosis and therapeutic decisions and ultimately enable better predication of clinical outcome of HGA patients.

3.4 METHODS

Time of survival (days or weeks) and subtype designations were available for 239 patients. The microarray expression matrix of this subset of patients was normalized as previously described. As an extension of DIRAC, a distance matrix was constructed for each selected network based on the pairwise orderings of the genes within the network. For example, if a network *m* consisted of six genes, there could be $\binom{6}{2}=15$ distinct ordered pairs; for a gene pair *i* and *j*, let X denote their corresponding expression values. If $X_i < X_j$ or $X_i > X_j$ for both patient A and B, the distance of these two patients was 0; otherwise the distance was 1. The direct sum of the distances for all 15 possible comparisons was then normalized by the size of the network, to give the final average distance of patients A and B on network *m*. We repeated this procedure for all patients to obtain a 239 × 239 distance matrix.

Unsupervised clustering in Matlab resulted in two groups (linkage method: weighted average distance, **Figure 3.2**). The first split on top of the dendrogram separated the samples into the two largest groups (A, B). If one of the groups (e.g., B) did not contain at least 10% of all samples, its samples were considered as outliers and removed from subsequent analysis; the remaining group A was then split into two (A1 and A2) according to the next joint on the dendrogram. Further outlier removal and group splitting continued until two groups with reasonable sizes were determined. The Kaplan-Meier method was used to estimate the survival distributions. Log-rank tests were used to evaluate the difference between survival groups. To address the issue of multiple hypotheses testing, stringent Bonferroni correction was applied to *P*-values obtained from log-rank tests.

3.5 CHAPTER 3 FIGURES AND TABLES

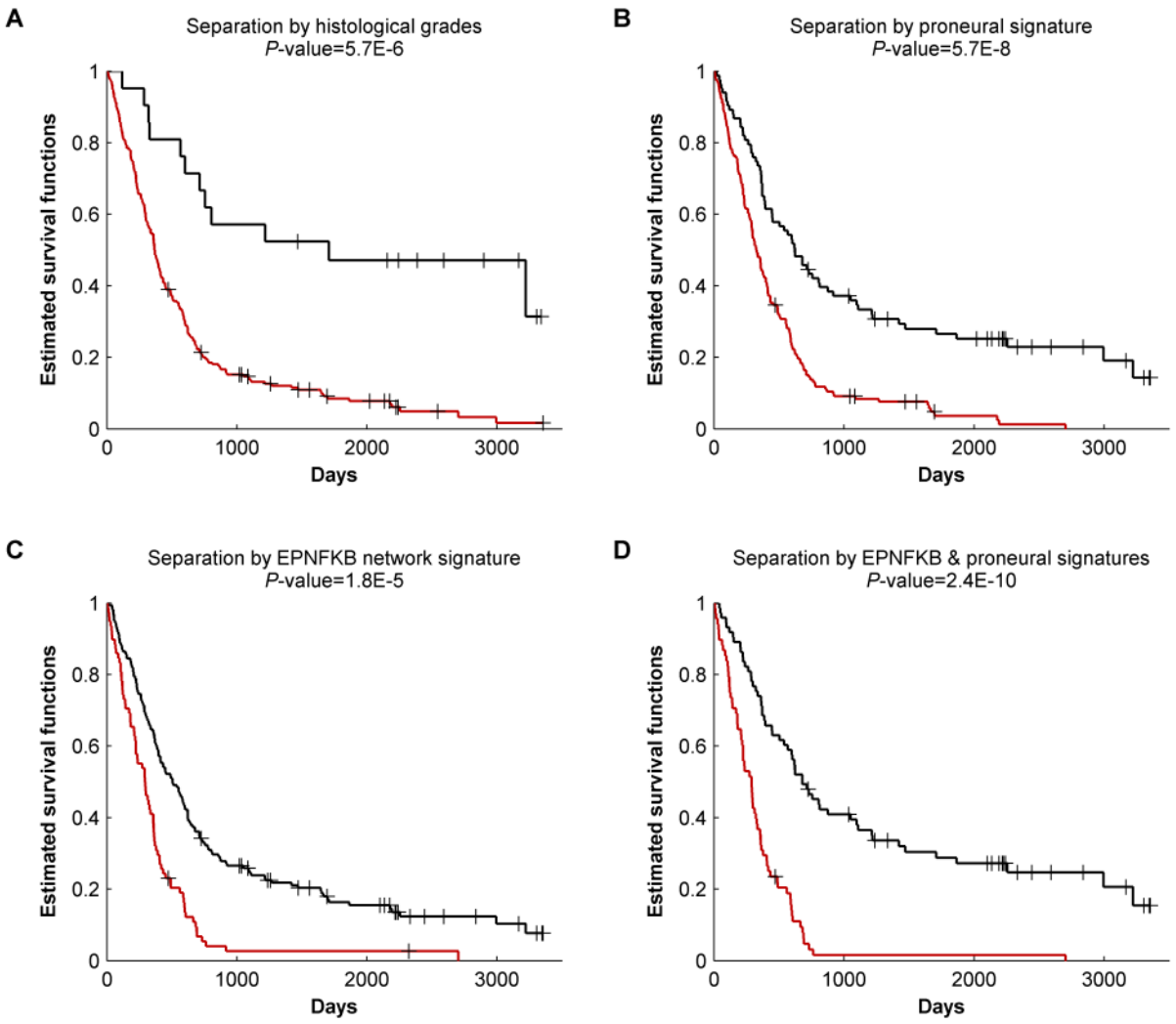


Figure 3.1 Comparison of different approaches to re-classify HGAs

Censored data indicate the patients being alive by the end of the study. a) Survival estimates of two groups separated by histological grades b) by *PN* vs. *non-PN* subtype, c) by prognostic network marker EPONFκB, and d) by combined status of *PN* (*non-PN*) and *EPOLONG* (or *EPOSHORT*). The log-rank test on the combined case has a more significant prognostic value than existing best classification.

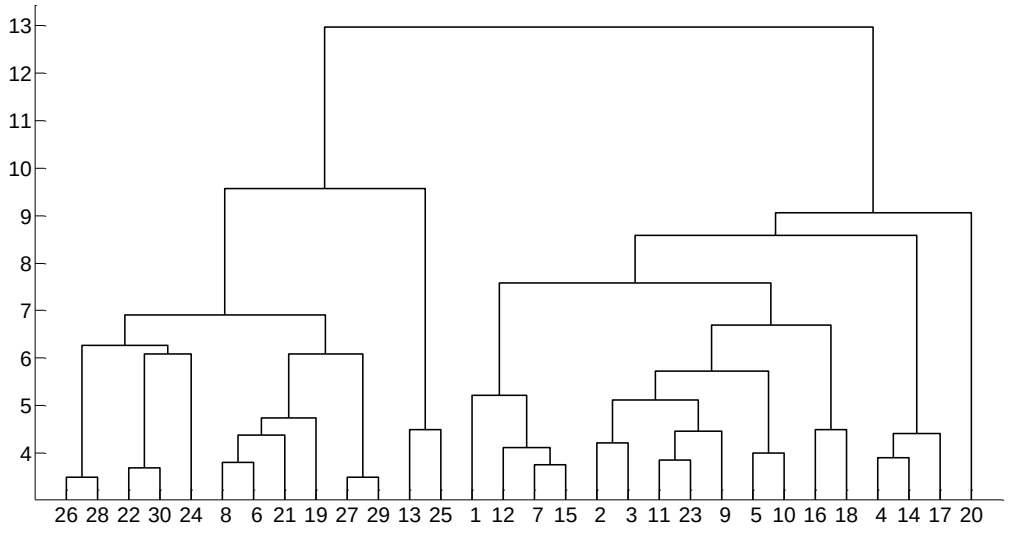


Figure 3.2 EPONFκB separates the patients into two clusters with survival difference, using hierarchical clustering

BioCarta Network	Survival Difference
EPONFκB	$1.8*10^{-5}$
CARDIACEGF	$6.8*10^{-3}$
IL22BP	$1.1*10^{-2}$
EPO	$1.1*10^{-2}$
FIBRINOLYSIS	$1.3*10^{-2}$

Table 3.1 Networks that can separate HGA patients into groups with statistically significant survival difference

CHAPTER 4 ANALYZING PROTEOMIC DATA FROM GENETICALLY ENGINEERED *MUS MUSCULUS* STRAINS

4.1 UTILIZING HIGH CONSISTENCY OF PROTEOMICS DATA

As I explained briefly in Chapter 1, though proteins are the translated products of RNA, protein concentrations do not necessarily correlate with the amount of corresponding mRNA in the cell [82]. Proteomic data obtained by mass spectrometry-based analysis and iTRAQ [83], Selective Reaction Monitoring (SRM) [84], and the newly emerging SWATH technologies [85] can provide direct quantification of thousands of proteins in the cell simultaneously.

The high consistency of protein measurement comes with a price, compared to gene expression data, which is intrinsically much noisier [86]. Protein measurement experiments are much more expensive and cover fewer targets compared to gene expression profiling. The most thorough and comprehensive protein experiments presently include only a few thousand proteins with fewer samples in per condition. The higher consistency, low noise level, and fewer measurement properties of proteomic data demand a unique processing and analysis pipeline that is fundamentally different from our approach to transcriptomic analysis in the last chapter. In this chapter, I focus on developing a framework that is statistically valid and sound, computationally simple and efficient and lastly, biologically valuable and meaningful.

As previously stated, GBM is the most common human brain tumor and it has a high fatality rate; the experiments conducted in this section are also aimed at understanding more on brain tumor biology and developing potentially new therapies based on these discoveries. We aimed to query and identify dynamically perturbed modules in a progressive mouse model of glioblastoma, by tracking behavior changes of regulatory networks in early, middle and late stages in disease development through protein measurements.

Most GBMs exhibit frequent aberrations in three prominent signaling networks: cyclin-dependent kinase/retinoblastoma (CDK/RB), receptor tyrosine kinase

(RTK)/RAS/PI3K/PTEN and MDM2/p53 [87]. Retinoblastoma RB is a tumor suppressor, and impaired RB drives tumor progression and is associated with short survival time [88]. On the other hand, the v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog (KRAS) protein is a member of Ras family, and activation of the KRAS gene is an essential step in cancer development [89]. Lastly, the phosphatase and tensin homolog (PTEN) also acts as a tumor suppressor gene whose loss of function leads to increased cell proliferation and reduced cell death [90].

The present work develops a methodology to process and analyze proteomics data collected from three mouse strains, which had been genetically engineered to contain one or more combinations of gene mutations in the three major networks mentioned above. Eventually, we hope to identify and validate potential protein targets that play significant roles in GBM initiation and progression, and discover potential drug targets and treatment strategies of GBM.

4.2 Overview of Experimental Design, Data Processing and Analysis

Three strains of mice, with two duplicates for each strain, were collected at four different time points: control (beginning of experiment), early, middle, and late. The strains were named TR-cre, TR-het, and TR-null, each of which had two, one, and no copies of PTEN, respectively (**Table 4.1**). Protein content was measured using iTRAQ for all four stages and for all three strains and a comprehensive set of more than 2000 proteins in most cases were quantified and recorded. The controls had no genetic mutations and served as reference channels for the tumor samples. Cancer samples at different time points were subsequently normalized against their references to obtain a relative expression level. The arithmetic average value of duplicate observations for the same stage was calculated to represent the overall protein content for that stage. After these steps of normalization and processing, we have 3 data points for each strain and 9 points for all strains.

Due to the extremely small size of samples for each condition, conventional gene analysis tools, such as Gene Set Enrichment Analysis (GSEA), which is a computational method to determine whether a set of genes show statistically significant differences between two

biological states, phenotypes and conditions are not suitable [91]. For the current study, it is almost impossible to reach statistical significance if we use the Kolmogorov-Smirnov test chosen by GSEA. We also cannot use Top-Scoring Pair (TSP) or DIRAC, both excellent methods to identify simple disease classifiers, for similar reasons. The unique nature of proteomics data calls for a method that is computationally very simple, and at the same time offers statistical power even if we only have one or two samples in each condition.

With these considerations in mind, I developed a method that groups genes into *a priori* defined sets of genes and calculates if the median of this particular set of genes show a difference between two cancer stages. By grouping genes into networks and calculating if the median expression value changes significantly as a group, the significance of networks could be evaluated using the Wilcoxon rank-sum test, which compares two matched samples to assess whether their population median ranks differ [92]. It is an alternative to the paired student's t-test, without the assumption that the populations are normally distributed [93]. The non-parametric nature of the Wilcoxon is especially suitable for the present data because we have no prior knowledge on the distributions of the genes in a gene set.

The manually curated gene set database I selected for data processing was the Biocarta pathway database, which was defined according to the BioCarta gene sets collection in the Molecular Signatures Database. Genes were grouped into 248 human signaling networks [91]. Before grouping gene expressions into BioCarta gene sets, we mapped mouse genes into their human orthologs according to Mouse Genome Informatics (MGI) database [94, 95]. Gene sets having at least two genes were kept in the collection for further analysis. Around 2000 genes were mapped to 159, 148, 157 gene sets in TR-cre, TR-het, and TR-null strains respectively.

4.3 EXAMINATION OF PERTURBED NETWORKS IN DIFFERENT STRAINS

We were interested in identifying networks displaying different group behavior, as measured by the network median under different cancer stages. To identify the networks, I compared expression profiles at different time points against each other (early vs. middle,

early vs. late, middle vs. late) and ranked the networks according to the *P-values* of rank-sum test. Networks that showed significant changes (P-values < 0.05) under various cancer stages were named *perturbed networks*. Perturbed networks could either show *up-regulation* if the network median increases, or *down-regulation* if the network median decreases. The results of up- or down-regulation varied significantly from strain to strain.

For all three strains, we used the symbol μ to represent the median value of the gene sets, and μ_{early}, μ_{mid}, μ_{late} to represent the corresponding network medians at different time points. **Figure 4.1** showed how different networks responded to cancer progression in the three cancer stages.

Both the TR-cre and TR-null strains (**Figure 4.1A and C**) demonstrated an overall pattern of up-regulation when GBM progressed from early to middle and late stages. Higher expression at more advanced tumor stages usually implies more genetic activity at the cellular level, and could reflect the body's counter efforts to restore the diseased cells to normal states. Genes involved in these correction mechanisms could be tumor suppressors or targets of tumor suppressors which act to prevent the tumor from advancing or promote immune responses in tumor [96-98]. The majority of genes being up-regulated might also be oncogenes or targets of oncogenes which are involved in tumorigenesis pathways and cause tumor cells to survive and proliferate [99-101]. Another possible reason for the observed differences in expression is how the raw data were normalized after initial measurement. However I was not able to access raw data to assess this hypothesis. .

In contrast to these two strains, the TR-het strain displayed much lower gene expression at the network level when the tumor progressed to the late stage (**Figure 4.1B**). The reason that most genes are under-expressed might be that the glioma had rapidly evolved to glioblastoma, and extensive necrosis had occurred in the late stage. The possible presence of large amount of cell death and microvascular hyperplasia might help to explain the low or even lack of activity of most genes [102, 103].

4.3.1 PERTURBED NETWORKS FROM THREE STRAINS

For the TR-cre strain, I observed a total of 10 networks for which the median expression value increased from early to middle stage, and from early to late stage (**Figure 4.2A**). I also identified another 9 networks with up-regulated average expression in one, but not both cases (i.e., up-regulation from early to middle, ***or*** from early to late stage). I did not discern any pathways with significant changes from middle to late stage.

The perturbed networks from The TR-het strain, in which one copy of the tumor suppressor PTEN was knocked out, demonstrated very different behavior from the TR-cre strain. There were 87 networks with $\mu_{mid} > \mu_{early}$, but no networks showed increased μ from early to late stage, or from middle to late stage. The top 10 perturbed networks ranked by *P-values* are shown in **Figure 4.2B**. In fact, all genes showed down-regulation advancing to late GBM. As explained earlier, the rates of progression in different strains vary significantly from each other, and in this strain, a high percentage of necrosis or apoptosis in cancer tissue might already be present in this terminal stage. The extremely low gene expression levels did not reflect meaningful biological events and it is also not valid to compare progression events at this stage to earlier stages, or to stages of other strains, therefore the particular set of data was not kept for further analysis.

Late stage GBM appeared earlier in the mouse strain with heterogeneous PTEN (PTEN +/-), but not in other two strains, which might be explained by the stochastic p53 mutation driven by KRAS activation. It has been hypothesized that KRAS activation would positively select for p53 missense mutant cells [104]. P53 tumor suppressor proteins are encoded by the TP53 genes, which guard and conserve genome stability by preventing genome mutation [105]. Gene mutations can change the resultant protein structure, resulting in effects on cell replication. Tumor cells that are genetically unstable may be allowed to replicate [106]. In fact, this gene represent one of the most frequently occurring perturbations in human tumors [107]. There are at least two genetic pathways leading to GBM, a *de novo* pathway without P53 mutation, and a progressive pathway with P53 mutation [108]. P53 mutations in gliomas coupled with heterozygous deletion of PTEN allele may imply an additive effect of

combined gene deficiency that drives glioma progression, therefore explaining the fast tumor growth in this case. Supporting this hypothesis is the finding that PTEN +/- mice with concomitant inactivation of CDKN1B, a key player in CDK/RB pathway, showed accelerated neoplastic transformation, and these mice all developed prostate cancer within the first three months of life [109].

The TR-null mouse strain, with both PTEN alleles inactivated, displayed a more similar gene and network pattern to the TR-cre strain, than to the TR-het strain, with middle and late protein profiles showing higher network medians than early stage protein profiles (**Figure 4.2C**). The experimental results suggest that homozygous deletions at the PTEN locus might slow the glioma development process compared to the heterozygous deletions. PTEN -/- is a much rarer event in gliomas than PTEN +/-; while PTEN LOH was detected in 30% primary GBMs [69, 110], only 5-10% loss of both PTEN alleles were detected in similar cases [111, 112]. Gliomas losing both alleles of PTEN might represent a distinct subset of GBMs that worth more effort to investigate the mechanisms of tumor initiation and progression [104].

4.3.2 SELECTION OF TARGETS FOR VALIDATION

After I examined all networks in each strain, over all stages, I concentrated my efforts to identify the most perturbed networks **appearing across multiple strains,** and subsequently selected a subset of genes that could be validated by Western blot to confirm their significance. I only selected networks for which the network median either increased or decreased in at least 1 out 3 possible scenarios ($\mu_{early} < \mu_{middle}$, $\mu_{early} < \mu_{late}$, $\mu_{middle} < \mu_{late}$). This yields a total of 268 gene targets across all strains. The numbers of commonly shared significant genes between any two strains and among all three strains are indicated on **Figure 4.3**. Among all the 268 genes, 18 were selected based on their known roles in GBM progression as well as *P-values* associated with the networks they came from. If a gene appeared in more than one network, *P-values* for that particular gene was averaged across all the networks in which it appeared (see **Chapter 4.4 detailed methodologies**). Western blotting for these proteins is currently ongoing.

4.4 CONCLUSIONS

In this chapter, I describe the development of a methodological and systematic way to process and analyze proteomics data, which reflect the amount of protein content present in cells. This project aimed to identify individual proteins and networks that were most perturbed when GBM developed as a result of mutations in one or more major signaling pathways, common to GBM. Protein profiling was done on three genetically engineered mouse strains and over 2000 proteins were tracked at three tumor development stages: early, middle and late. The TR-het strain, in which one copy of PTEN was knocked out, displayed faster tumor progression than TR-cre, in which both PTEN copies were kept intact and TR-null, which lost both copies of PTEN. I aggregated related genes into gene sets and analyzed their collective behavior, captured by a single metric called *network median* (or μ) across stages, and assigned statistical significance based on Wilcoxon rank-sum test. The most perturbed networks ranked by *P-values* from each strain were identified. A small subset of genes selected from these significant networks across all strains is now waiting for further experimental validation to confirm their potential values in GBM initiation and progression.

4.5 DETAILED METHODOLOGIES

For each network with more than two genes, a network median across all gene components in the network was computed at each time point in the dynamic time series data (i.e., early, middle or late). For example, if a network consists of three genes, and they have expression values 1, 2, 3 at early stage, and the same set of genes are expressed at 5, 7, and 9 at middle stage, we will compare the early median value 2 with the middle median value 7. The statistical question set out to answer is: Is 7 significantly higher than 2 given the observations? The Wilcoxon rank-sum test was used to evaluate the significance and direction of change.

If the question is framed in statistical language, we tested the null hypothesis (H_0), which states that the median difference of this particular gene set under two conditions is zero,

against the alternative hypothesis (H_1) that the median difference is not zero, and whether the median increases or decreases if the alternative hypothesis is true.

For each gene expression value, $|X_{2,i} - X_{1,i}|$ was calculated and the signs of the differences were noted. Pairs with zero differences ($|X_{2,i} - X_{1,i}|=0$) were excluded to reduce the sample size. Next, the absolute differences were ordered from largest to smallest and their ranks (R_i) in the ordering were recorded. In our example, the three absolute differences 4, 5 and 6 received a rank of 1, 2 and 3 respectively. Pairs with the same ties received a rank equal to the average of their ranks.

The test statistic was calculated as follows:

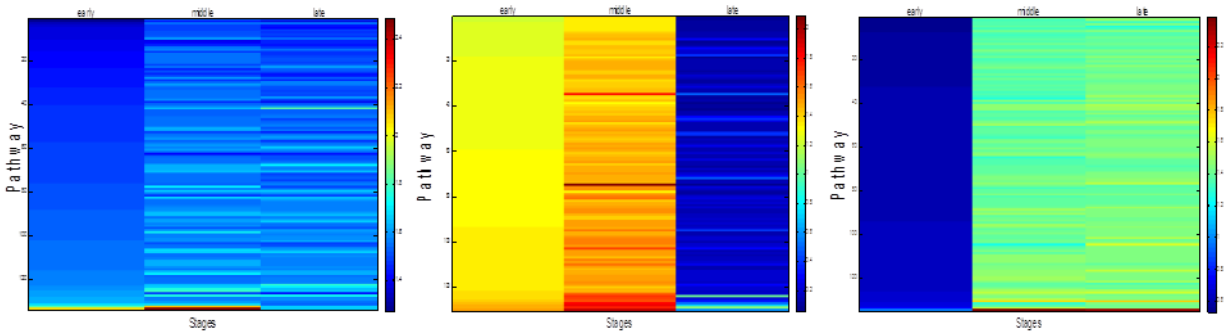
$T = \sum_{i=1}^{N} sgn\left[\left(X_{2,i} - X_{1,i}\right) * R_i\right]$.

Last, T was compared with cutoff value to decide whether to keep or reject the null hypothesis. If the alternative hypothesis is accepted, i.e if the comparison is statistically significant, a score of 1 would be assigned to this test (i.e., middle vs. early). Similar procedures were repeated for early vs. late, and middle vs. late; scores are assigned based on the statistical test results. The maximum possible score for a network is 3, and this happens when both middle and late stages showed significantly higher network medians than early stage, and at the same time, the metric was also higher in late stage than in middle stage.

After the sum of scores (possible values 0, 1, 2, 3) were calculated for all networks, the networks were ranked according to the sum of scores. For networks with the same score, *P-value*s were used to break the ties and gave them different significant levels.

The genes in the statistically significant networks were aggregated for each strain; I found that a large amount of genes were shared among different strains. These genes showed their essential roles in GBM independent of genetic backgrounds. To make sure that all three strains were contributing evenly to a 250-gene list, all networks in the TR-cre and TR-het with scores 1 and above, and networks in the TR-null strain with scores at least 2 were selected for further examination. .

4.6 CHAPTER 4 FIGURES AND TABLES



A B C

Figure 4.1 Heatmaps of network medians at different GBM stages

A) TR-cre strain B) TR- het strain and C) TR-null strain. Network medians are ranked from low to high in their respective early stage

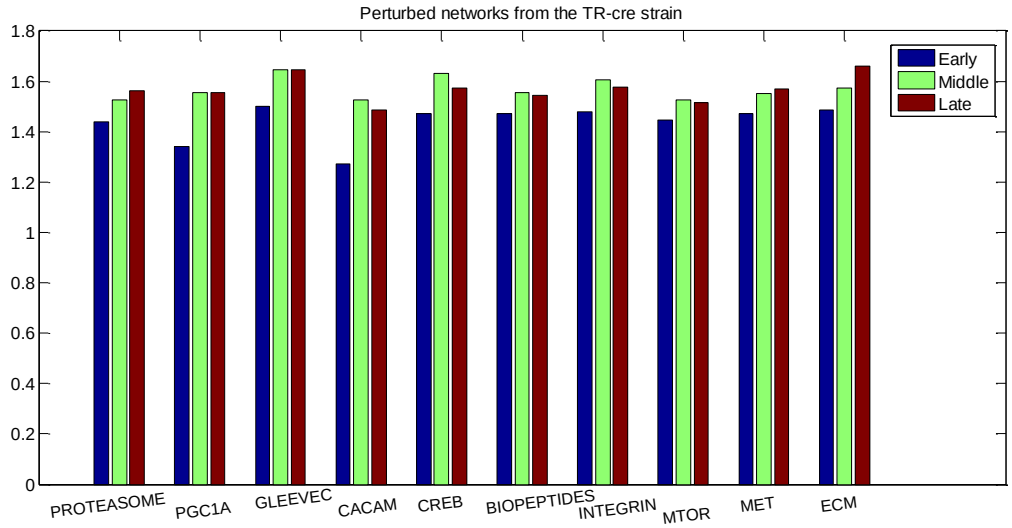


Figure 4.2A Most perturbed networks from the TR-cre strain

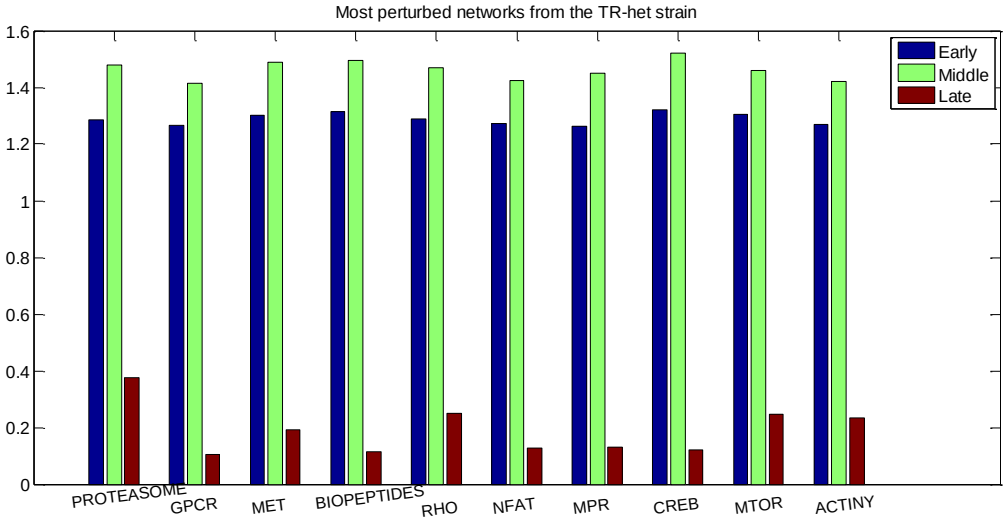


FIGURE 4.2B Most perturbed networks from the TR-het strain

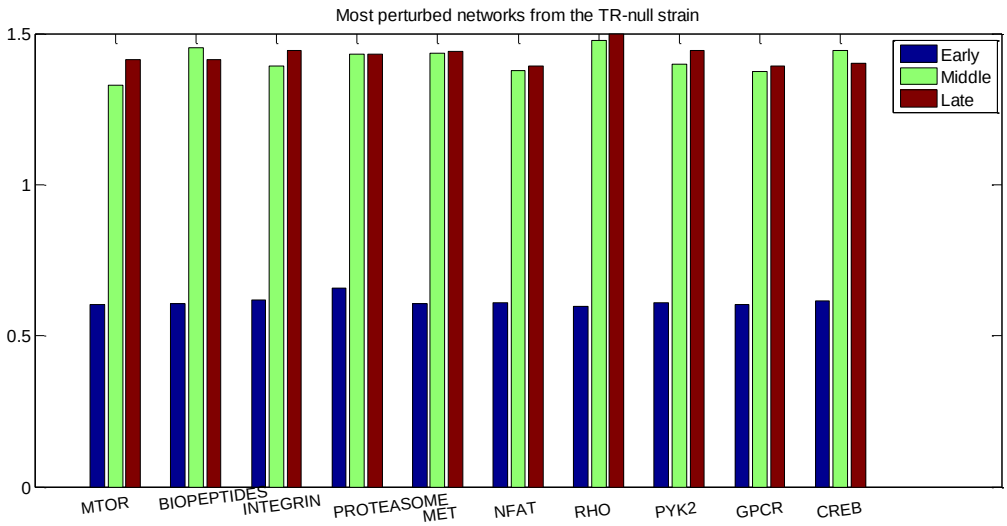


FIGURE 4.2C Most perturbed networks from the TR-null strain

The vertical axis indicates relative average network median μ after normalized as percentage of the control channel, and the horizontal axis lists the names of the networks. Early stage is shown as blue, middle as green and late as red bars. Most perturbed networks showed up-regulation in more advanced cancer stages.

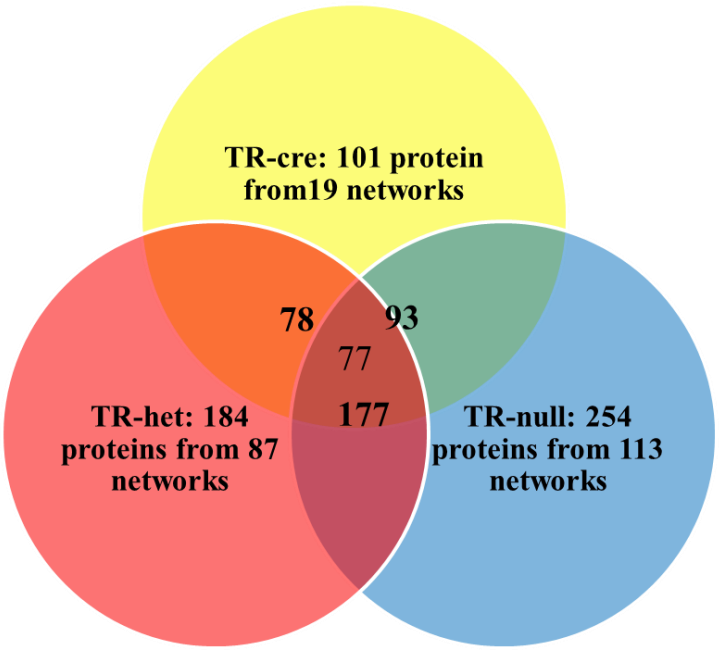


Figure 4.3 Gene sets selected from perturbed networks

The three circles represent perturbed networks and the gene sets in them. The three numbers 78, 93, and 177 represent common genes shared by two adjacent strains, and 77 genes are shared across all three strains. The numbers are not proportional to the area in the Venn diagram.

	RB	KRAS	PTEN
TR-cre	inactivated	activated	+/+
TR-het	inactivated	activated	+/-
TR-null	inactivated	activated	-/-

Table 4.1 Combinations of gene knockouts for different mouse strains

CHAPTER 5 RECONSTRUCTING LIVER METABOLIC MODEL FOR *MUS MUSCULUS*

In this chapter, I describe my work in building the first liver metabolic models for mouse. In order to reconstruct metabolic models in a tissue-specific context, we need to start with a functional generic model, which contains information for all reactions and metabolites in a generic cell. From there, we look for evidence of which genes and reactions exist in a liver cell and remove reactions without such evidence. Therefore this chapter is divided into two parts: first, examine and refine genome-scale models for mouse, and second, build liver models out of the generic mouse model.

5.1 REFINING GENOME-SCALE METABOLIC MODELS FOR *MUS MUSCULUS*

5.1.1 RECONSTRUCTION OF GENOME-SCALE METABOLIC MODELS

As I explained in the introduction, metabolic analysis can offer a distinct perspective and valuable insights into the molecular mechanisms of a particular organism. A *metabolic model* is a mathematical representation of the biochemical pathways in the metabolic network of an organism. Each reaction consists of metabolites, genes, transcripts, proteins and the network consists of inter-connected reactions. For a certain organism, physiochemical constraints (e.g. conservation of mass) are added after the metabolic network has been assembled. A stoichiometric matrix S, of size $m \times n$, where m is the number of participating metabolites and n is the number of corresponding reactions in the network, is used to describe the relationships between metabolites and reactions. Each element $S_{i,j}$ in the matrix represents the stoichiometric coefficient of the metabolite in the corresponding reaction. Constraint-based modeling typically analyzes metabolic fluxes at steady state ($S \times v = 0$, where v is the flux distribution vector containing flux values for each reaction)—i.e. zero net change in metabolite concentrations. Other constraints such as upper and lower reaction bounds (maximum flux v_{max} and minimum flux v_{min}) and reaction reversibility, when known, are placed on each reaction. The resulting model is then ready to be used for phenotype simulations using various constraint-based reconstruction and

analysis (COBRA) methods [113]. When the metabolic biochemical pathways are integrated with whole genome sequences, a *genome-scale metabolic model* is constructed.

The first genome-scale reconstruction of metabolic networks, which appeared in 1999, was for *Haemophilus influenzae Rd* [114]. Since then, different methods for building and analyzing metabolic networks of genome-scale models have been established for all branches of life [115, 116].

Because disease results from the malfunction of one or more biological networks, diseased states show aberrant metabolic networks as compared to the normal state. Metabolic genes controlling reactions in the networks exhibiting abnormal expression levels can be identified in gene expression data. Tracking the activities of these significant genes or biomarkers can help us distinguish diseased vs. normal states. Therefore, metabolic reconstruction is an effective approach to infer biomarkers that may represent critical nodes in the perturbed networks, leading to insights into new drug treatment targets. These new therapies could either be used to convert the diseased network back to a normal state or permit the specific killing of the diseased cells [117].

5.1.2 METABOLIC RECONSTRUCTIONS FOR MUS MUSCULUS AND HUMAN

The mouse serves as a fundamental experimental animal to mimic human diseases to improve understanding of the causes and progression of disease symptoms [118-120]. In this chapter, I examine the existing efforts to reconstruct genome-scale mouse metabolic models and made necessary changes to improve and refine its functionality.

The largest and most comprehensive mouse metabolic reconstruction to date contains 1,415 metabolic genes accounting for reactions in eight cellular compartments (cytosol, mitochondrial, extracellular, golgi, lysosome, ribosome, and nucleus) [120]. Despite these capabilities, the current model lacks the ability to simulate several essential metabolic functions. For instance, simulations of the current mouse model failed to produce any of the nine non-essential (should be synthesized *de novo* by mouse) amino acids (AAs) when provided with glucose and other inorganic metabolites. Synthesis of non-essential AAs within the body without supply in the diet represents a fundamental ability of mammalian

metabolism and the inability of the model to complete this implicates missing reactions in major metabolic pathways.

Dealing with these limitations, missing links in this mouse metabolic network model had to be identified. The improved model needs to pass a universal functional test, which includes functional central metabolic pathways (glycolysis, TCA cycle, pentose phosphate pathway), a functional fatty acid synthesis pathway (from acetyl-CoA to palmital-CoA), and demonstrates the capability of synthesizing non-essential amino acids, nucleotides and key membrane lipids from glucose.

Released in 2007, *Homo sapiens* Recon 1 is a comprehensive literature-based genome-scale metabolic model that accounts for the functions of 1496 genes, 2766 metabolites, and 3311 reactions. [121]. In 2013, Recon 2 which represents a "consensus metabolic reconstruction" and the most comprehensive representation of human metabolism was released. Compared to its predecessors, the reconstruction has improved its functional and topological features, doubled its reaction numbers and included many more unique metabolites [122]. Due to the high sequence homology between human and mouse (range around 85%~92%), a mouse metabolic model was published in 2010; the model was built by searching for genes homologous to Human Recon 1 within the mammalian genome [120]. The draft model contained 1,415 metabolic genes and was termed (***iMM1415)***. This model represents the largest and most comprehensive mouse reconstruction to date. Unlike the corresponding human model, there had been limited efforts extending this mouse metabolic model to better our understanding of mouse metabolism. Specifically, there has been no study focusing on building tissue-specific models for the mouse. Moreover, the current mouse models still lacks some fundamental metabolic capabilities as a generic model (**Table 5.1**).

5.1.3 IDENTIFYING LIMITATIONS AND INACCURACIES IN THE GENERIC MODEL

I obtained an updated version of the generic mouse model from the author of *i*MM1415. This revised model consists of 1752 transcripts representing 1361 unique genes (*i*MM1361). The reduction in the number of genes compared to *i*MM1415 is due to removal of some genes being human genes in the former model.

A universal metabolic test was used to determine whether *i*MM1361 possesses the basic functionality of a cell. Metabolites tested include precursor metabolites in central metabolic pathways such as glycolysis and the TCA cycle, non-essential amino acids (AA), nucleotides, palmital-CoA, cholesterol, and several membrane lipids. This model is incapable of producing any of the non-essential amino acids and the nucleotides and certain lipids (**Table 5.1**). The production of non-essential amino acids were tested in the following condition: the networks were allowed to uptake glucose and inorganic compounds such as oxygen, carbon dioxide, phosphate, and other ions. All other organic compounds besides glucose were constrained to be efflux only. The production of nucleotides was tested in the same medium.

Linear programming approaches (flux balance analysis (FBA) or flux variability analysis (FVA)) in the COBRA toolbox were used to solve the equation $S \times v = 0$ given upper and lower bounds on the metabolite concentrations. FBA gives a particular flux distribution, while FVA gives flux boundaries of the reactions which could equally return equivalently optimal solutions to maximize or minimize the objective function [113]. In each test, the objective function was to maximize the production of the selected metabolite (i.e., any of the AA or nucleotides).

I traced the amino acid formation pathways and found two missing reactions involved in glutamate metabolism (BiGG ID 2169419, 2169428). These two reversible reactions synthesize glutamate from α-ketoglutarate with glutamate dehydrogenase. Two metabolic genes (EC number 1.4.1.2 and 1.4.1.4), have been added to the model accordingly. The existence of the reactions and related genes has been validated in literature [123, 124]. With the addition of these two reactions, the model is capable of producing all non-essential AAs, nucleotides and necessary lipids.

5.2 RECONSTRUCTION OF TISSUE-SPECIFIC MODELS FOR *MUS MUSCULUS*

5.2.1 NECESSITY TO BUILD TISSUE-SPECIFIC MODELS

Complex multicellular organisms such as mouse or humans consist of many distinct tissues and cell types, each only expressing a *fraction* of the metabolic genes encoded within the genome [121]. To accurately track changes in active genes for specific tissues involved in the initiation and progression of certain diseases, we need to study these diseases in a *tissue/organ-specific* context. Previous efforts have reconstructed and analyzed individual tissue models [125-128] for humans, while a few others have furthered the efforts by simulating the metabolism of a larger system containing multiple tissues or cell types and taking into account the interactions among them [129, 130]. Specifically, Recon 1 was tailored to describe metabolism in three human cells: adipocytes, hepatocytes, and myocytes. These three cell-specific networks were integrated using a novel multi-tissue type modeling approach to simulate known metabolic cycles and study diabetes [131].

Though a lot of efforts have been devoted to build tissue-specific models for humans based on human recon 1 and recon 2, there had been very few studies focused on building analogous tissue specific models for mouse. Our study aimed to create the first liver specific mouse models based on our improved and refined version of the generic model, presented in Chapter 5. The main focus for this study is the liver, for which the major metabolic cell type is the hepatocyte. The liver represents an essential metabolic organ in which glucose circulates through after it enters the blood; it uses alternative sugars as energy sources, completes the urea cycle, and produces urea from nitrogen. Many metabolic disorders and diseases such obesity, diabetics, and fatty liver diseases all involve the crucial body organ liver. In this chapter of my dissertation, I present the reconstruction of two liver-specific metabolic models for mouse, with one for the normal (control) strain and one for mouse strains with diabetes. These two models were compared physiologically to infer metabolic pathways that were most impacted by the onset of diabetes.

5.2.2 *DIABETES AND DIET-INDUCED OBESITY IN HUMANS*

Diabetes mellitus is a chronic disease that typically requires intensive, lifelong management. According to the International Diabetes Federation as of 2011, 336 million people worldwide have type 2 diabetes, resulting in 4.6 million deaths each year, or one death every seven seconds [132]. In the United States, 12% of American adults, and >25% of those over the age of 65, are affected. Currently, there is no cure for diabetes. What exacerbates the problem is that diabetes increases the risk of heart disease, stroke, and microvascular complications such as blindness, renal failure, and peripheral neuropathy [132]. Type 2 diabetes mellitus (T2DM) makes up most of diabetic cases. Reported causes to this disease include lifestyle factors such as age, pregnancy, obesity, and genetic factors.

Over the past 50 years, we have witnessed a dramatic increase in T2DM. The massive increase in diabetes incidence is not due to genetic changes; rather, it is largely a consequence of the concomitant increase in obesity [133]. It is widely recognized that defective insulin secretion caused by reduced β cell function is the key problem in T2DM. Insulin transports glucose entering the blood stream to the muscle, fat, and liver cells—where it can be used as energy sources to the body. Obesity lowers insulin sensitivity in peripheral tissues; then, to compensate for this, β cells up-regulate insulin secretion. The degree to which they are able to do so determines whether or not the individual develops diabetes [134]. Obesity and related T2DM are generally accepted as a consequence of dietary imbalance rather than genetically programmed diseases [135]. They are modulated by lifestyle and diet, which induce pathophysiological changes throughout the body [136]. Several genes implicated in T2DM have also been identified, but how exactly they interact with each other remains enigmatic.

5.2.3 *GENETIC STRAINS OF OBESITY AND DIABETES OF MOUSE*

As I introduced earlier in Chapter 5, the mouse is a primary mammalian model system for genetic research. With available inbred knockout mouse strains, the mouse metabolic reconstructions could be examined for their phenotype prediction capabilities [137]. It is

essential that the rodent models imitate particular characteristics of human disease and resemble the genetic changes in diabetic and obese patients.

Established genetic strains of obesity and diabetes include db/db mice, ob/ob mice, Zucker diabetic fatty rats, etc. [135]. These models are obtained by inducing mutations in certain chromosomes. As previously mentioned, human diabetes and obesity are largely diet-induced, chronic consumption of a high-carbohydrate, high-fat diet by normal rodents provides an adequate rodent model to study these diseases of interest [135].

A common inbred strain of laboratory mice, which is usually used as a background for these genetic variants is C57BL/6 ((often referred to as “C57 black 6” or “black 6”). Its popularity is largely due to the availability of congenic strains, easy breeding, and robustness. Two distinct substrains of the B6 mice, C57BL/6J from Jackson Laboratories ("J") and C57BL/6N from NIH ("N") were later developed, distributed, and maintained by different investigators [138, 139]. Though externally similar, these two sub-strains are genetically distinct from each other. I built reconstructions exclusively for the C57BL/6J strain, though the protocol could be easily extended to other mouse strains, if needed later. Subsequent analysis of these models will provide us insights into the pathophysiology of the disease (e.g., diabetes) and selection of the potential therapeutic targets.

5.2.4 ALGORITHMS FOR AUTOMATIC RECONSTRUCTION OF TISSUE-SPECIFIC MODELS

Earlier in this chapter, I presented my improvement of the generic mouse model, which laid a solid foundation on which to build the liver-specific reconstructions. Having picked the mouse strain (B6) and decided on the disease to study (diabetes), the next step was to select an algorithm, a computational method to decide which reactions from the generic, whole body model should stay in the final liver model, and which should not, based on gene expression data that tell us which genes are expressed in the tissues.

The model-building algorithm (MBA) by Jerby *et al.* addresses this challenge [128]. It derives a tissue-specific metabolic model from a generic one based on network integration with various molecular data sources. First, core reactions are inferred from gene expression data. They are further split into two groups: reactions with high and moderately high

likelihood reactions (C_H and C_M). The final optimal model will include all C_H reactions, a maximal number of C_M, and a set of gap filling reactions.

To determine whether to remove or keep a certain reaction from a given iteration, the potential reactions were scanned in random order. The scanning order of candidate reactions affects the resulting model. Therefore, the algorithm is executed 1000 times with different, random pruning orders to construct multiple candidate models. These candidate models were further compared and analyzed to reach the final viable and consistent model.

However, the accuracy of the optimal model is limited by the fact that even 1000 iterations could only cover a small portion of the large space of possible orderings. Another problem with MBA is its modeling building process is very time-consuming. A more deterministic and simulation-independent ranking of potentially removable reactions could help to accelerate model building time dramatically [140]. Our laboratory recently developed a method called metabolic Context-specificity Assessed by Deterministic Reaction Evaluation (mCADRE) (**Figure 5.1**), which fulfills the same purpose as MBA, but with two additional advantages: first, non-core reactions are ranked according to their own expression evidence and connectivity map to other reactions in the network and then removed in the inverse order of this ranking [140]; second, the performance of mCADRE was significantly better than MBA. Its performance was evaluated by reconstructing a human liver model and comparing it with the model built using MBA. mCADRE demonstrated improved model functionality, and significant shorter reconstruction time (under the same configuration, mCADRE required only about 10 CPU-hours while MBA took approximately 10,000 CPU-hours) [140].

5.2.5 DATA COLLECTION AND PROCESSING FOR BUILDING TISSUE-SPECIFIC MOUSE MODEL

Microarray gene expression data were used to identify highly expressed or non-expressed genes for both normal and disease models. For each model, I compiled raw microarray CEL files from previous studies as cataloged in the NCBI Gene Expression Omnibus (GEO). I focused strictly on data from the microarray platform Mouse Genome 430 2.0 (Affymetrix,

Santa Clara, CA). **Table 5.2** lists the number of samples and studies collected for normal and diabetic liver. For the control tissue model (**Table 5.2A**), I included only gene expression data for the B6 mouse strain. Within this strain, I excluded gene expression profiles with genetic modifications on the B6 genotype (transgenic or knockout models, models with gene mutations etc., because these changes may directly affect their phenotypes (e.g., weight gain), thus defeating our purpose of building a "control" strain for mouse). For the disease (i.e., diabetes) model (**Table 5.2B**), I only included expression obtained from F2 mice inbred from a diabetic strain and B6.

In order to read the probe sequences from the Affymetrix platform (i.e., Mouse Genome 430 2.0), I obtained two relevant files: first, a FASTA file containing the probe sequence information [141], and second, the CDF library file which specifies which probe set each probe belongs to on the selected GeneChip array [142]. With these two files ready, I used the "affyprobeseqread" command in Matlab Bioinformatics Toolbox to obtain the structure containing the probe set IDs from the selected mouse platform.

The "consensus pre-processing" method which I used in Chapter 2, was again applied to process CEL files to normalize differences introduced by non-uniform studies and sample preparation procedures. This method is described in greater detail in [30]. This algorithm returned us with normalized gene expression files, and the associated Presence/Absence calls and the corresponding probes for each gene feature.

The metabolic mouse reconstruction identifies genes by Entrez Gene IDs, rather than by probe set IDs from consensus pre-processing. It was necessary to find the correspondence between two sets of identifiers and filter out probes without matched gene IDs. The BioMart software, which offers the free service of converting between different formats of gene identifiers, was used to complete the gene mapping task [143]. I started with 45101 probe sets on the selected mouse platform and ended with 20963 genes after these steps.

5.2.6 GENERATION OF TISSUE SPECIFIC MODELS USING MCADRE

As explained earlier, mCADRE is a very efficient and robust automatic reconstruction algorithm that prunes reactions with insufficient evidence to stay in the tissue models. A few

key metrics were calculated and tuned before and during the pruning process, which influenced heavily the final reactions included in the tissue model. Before all potential reactions could be lined up for examination, they needed to be *ranked properly*.

For each gene, a ubiquity score *U(g)* was calculated by quantifying how often a gene *g* is expressed across all samples of the tissue of interest. A ubiquity score ranges from 0 to 1 and is a parameter indicating the prevalence of the gene in the tissue.

The expression-based evidence E_x was calculated for each gene-associated reaction by assigning specific rules to combine and integrate the ubiquity scores of multiple genes involved in the reaction [140]. E_x was used to rank reactions from high to low, and reactions with sufficiently high E_x ($E_x >= 0.9$ in my study) were defined as the core reaction set. For non-core reactions with low expression-based evidence or reactions without associations to known metabolic genes, a network topology metric connectivity-based evidence E_c was introduced to rank them. A third reaction ranking metric, confidence level scores E_l which provides literature support to each reaction in the generic mouse model, was added to further distinguish reactions with similar expression and connectivity-based scores. The details and formulae of ranking reactions followed from the original mCADRE paper and were described in detail in [140].

After the reactions were ranked, they were examined in order to determine whether they should be kept in the tissue model. Non-core reactions could be pruned from the generic model, if their removal affected fluxes through the core reaction set, or resulted in failure to produce key metabolites from glucose. A list of key metabolites was compiled based on literature evidence and included principle metabolites common to all cellular models [131]. These metabolites included precursor metabolites from central metabolic pathways (glycolysis, TCA cycle, pentose phosphate pathway), amino acids, lipids, and nucleotides. Successful production of these products from glucose represents the most essential metabolic functions that are common to both the generic model and to the tissue model. Therefore, this basic functionality test was carried out in the generic model before the pruning started, and was also examined for each non-core reaction, to make sure the removal of the candidate reaction would not result in the failure of this test.

For core-reactions, I followed suggestions from the original mCADRE paper to allow for a flexible core reaction set, which would increase tissue-specificity of metabolic pathways. The principle behind a flexible core is that, when there is enough strong evidence saying a reaction should not be included, even a core reaction could possibly be removed. The parameter representing the ratio of inactivated core reactions to inactivated non-core reactions was tuned to balance sensitivity and specificity: a low ratio keeps more reactions with strong positive evidence, while a high ratio removes more reactions with strong negative evidence [140]. This parameter was selected to be 0.2 for both the control and disease liver models. For the control model, if the ratio was increased to higher values (from 0.2-0.5 to 0.5-0.8), another 18 core reactions related to chondrointin/heparan sulfate biosynthesis would be excluded from the model, however, there is literature evidence stating these reactions should be present in mouse liver [144]. For the disease model, increasing the ratio would also exclude partial heparin biosynthesis pathway, so 1/5, which means for each core reaction to be removed, at least five non-core reactions need to be removed at the same time, was chosen for both models as the optimal balance to keep reactions with strong positive evidence, and exclude reactions with strong negative evidence.

After these parameters were experimentally determined, mCADRE was used to create the preliminary versions of *normal-liver* and *diabetes-liver*.

5.2.7 ADDING FUNCTIONALITY TO IMPROVE SPECIFICITY

The models coming straight out of mCADRE only possess the universal functionality for all cell types; they lack the specific characteristics unique to each cell type. For example, the hepatocytes are capable of generating glucose from various non-carbohydrate carbon substrates such as pyruvate, lactate and glycerol, through a process called gluconeogenesis. I collected a comprehensive set of liver specific functionality tests from various research studies and implemented them in both liver models.

Gluconeogenesis Tests

The hepatocyte is responsible for a wide range of biochemical functions and *gluconeogenesis* is amongst the most important responsibilities of hepatocytes. It is the

primary mechanism to maintain blood glucose level in mammals. The main gluconeogenic precursors are lactate, glycerol and two glucogenic amino acids (alanine and glutamine). These four precursors account for more than 90% of the overall gluconeogenesis [145]. Other less important gluconeogenic substrates include additional 11 glucogenic amino acids and 5 amino acids that are both glucogenic and ketognic. Another key substrate is pyruvate, which interconverts with lactate in the Cori Cycle and is the first designated substrate of the gluconeogenic pathway. The hepatocyte network was tested for its ability to produce glucose from all of above-mentioned substrates. Gluconeogenic simulations with the reactions, substrates and maximized glucose production were listed in **Table 5.3**. It is worth mentioning that metabolic simulations initially showed that only 19/21 substrates could generate glucose. The two substrates threonine (Thr-L) and methionine (Met-L) failed to produce glucose when they were provided to the network. We further examined the original generic metabolic network and similar problems also existed there. The problem was traced to the need to transport propanoyl-CoA, which is produced in the *cytosol* by an intermediate product 2-Oxobutanoate in threonine and methionine degradation, to *mitochondria*, so propanoyl-CoA could be used by downstream reactions to produce other essential intermediates in the metabolism pathway. By identifying the gap and completing the pathway, the addition of missing reactions further refined the metabolic network in the *generic* mouse model.

Ketogenesis tests

Another key function of hepatocytes is their capability to produce ketone bodies (acetoacetate and β-hydroxybutyrate) in the mitochondria in response to low glucose concentration in the blood. Carbohydrates are usually the first sources for energy, but when carbohydrate stores are exhausted, cells turn to fatty acids to generate energy. Ketone bodies are produced from the β-oxidation of fatty acids. The production of acetoacetate and β-hydroxybutyrate were tested using Stearoyl-CoA (C18:0) as a model fatty acid. The simulations can be found in **Table 5.4A**.

Alternative sugar metabolism tests

The liver is able to utilize alternative sugars other than glucose as energy sources [146]. I tested the ability of the hypatocyte networks to produce ATP from fructose, galactose and mannose. These three sugars were chosen to demonstrate the functionality of the network because all of them were metabolized primarily in the liver [147-149]. The simulations showed that the alternative sugar sources could produce the same amount of ATP on a molar basis as glucose (22.8 ATP/mol). The simulations can be found in **Table 5.4B**.

Amino acid degradation, ammonia and ethanol detoxification tests

Three degradation simulations were also tested on the liver metabolic networks: ammonia detoxification, ethanol degradation, and amino acid degradation. Ammonia is a by-product of amino acid metabolism. The amino group is removed and converted to ammonia-NH_3, which is a toxic compound and is converted to urea in the urea cycle. Eventually urea is eliminated from the body through the kidneys. I simulated the ammonia detoxification process by allowing for glucose uptake, optimizing for ammonia uptake, and the both liver models were able to convert ammonia into urea. Alcohol was metabolized first to acetaldehyde and then to acetate, which enters the Krebs cycle and eventually broken down into carbon dioxide and water. Similar to alcohol, amino acids could also be oxidized to CO_2 and H_2O to generate energy. The energy produced from amino acid accounts for 10 to 15% of all energy in the whole body [150]. The amino acids were first converted to ammonia and then enter the urea cycle, where they are converted to urea for excretion. Therefore, amino acid metabolism is intimately intertwined with ammonia and ethanol degradation. The simulation details of these three catabolic processes are listed in **Table 5.4C.**

Glycogen and cholesterol production tests

The last set of functionality tests is glycogen and cholesterol production. In liver cells, extra glucose is mainly stored as glycogen and functions as one important form of long-term energy storage. The inter-conversion between glucose and glycogen is as follows: when the body needs energy, glycogen is rapidly broken down into glucose-6-P and then enters glycolysis pathway, providing the cell with source of energy. When the glucose level in blood is high (e.g., after a carbohydrate-containing meal), glycogen synthesis is activated to

store extra glucose. The liver also synthesizes large quantities of cholesterol, which is an essential component of lipid membranes. Glycogen and cholesterol production from glucose were tested in the liver models in a manner similar to the precursor tests in **Chapter 6.6** and listed in **Table 5.4D**. The hepatocyte network could not synthesize glycogen from glucose at the beginning. The problem was traced to a missing transport reaction, which is responsible to bring Tyr-ggn (a primer for glycogen synthesis) from outside of the cell to the inside.

A total of six reactions were added to both liver models in an attempt to improve tissue specificity. The reactions and their associated genes and simulations are summarized in **Table 5.5**.

5.2.8 MANUAL CURATION OF LIVER SPECIFIC BIOMASS FUNCTION

The biomass reaction for the liver models was manually curated based on the published molecular content of the hepatocytes. The relative content of protein, lipids, carbohydrate, glycogen, water, and nucleotides is listed in **Table 5.6 and Figure 5.2**. As expected, the liver cells contain a large portion of protein (52%) because their role in synthesis and storage of proteins. Either mouse or rat was used to define the proportion of each nutrient whenever they were available; in the absence of direct experiment measurements collected for mouse or rat liver, published data for human was used to define the cell composition.

The protein proportion in the cell consists of 20 amino acids. The total amino acid content in the liver cell is further broken down in **Table 5.7A**. The values collected are measured in nmol per gram of liver tissue. In order to obtain the coefficients for the biomass equation, the values went through sequential calculations and eventually were converted to mmol per gram of dry weight of cell tissue.

The lipid distribution of the mouse liver cell including phospholipid, cholesterol, and triglycerols were also curated from different studies for mouse or rat (**Table 5.7B**). The nucleotide composition of mouse liver cells was adapted from reported values measured for generic mouse cells (**Table 5.7C**).

After the molecular content of a liver cell was properly defined according to published data, the generic biomass function was replaced by the corresponding tissue-specific biomass

equation to reflect the objective of a hepatocyte cell. I assumed that the disease cell and the normal cell have identical objectives to utilize the nutrients to grow and reproduce. In both types of cells, they were given a finite amount of glucose, essential amino acids, and fatty acids. Inorganic compounds and ions were not constrained. The relative amounts of each nutrient were largely adopted from [151]. After setting the biomass functions in both models as the objective function and maximizing for biomass production, a normal cell and a diabetic cell were found to have a maximized biomass flux of 0.0907 and 0.0977 mmol/g dry cellular weight (DCW) respectively.

5.2.9 DISEASE MODEL VS. CONTROL MODEL

After implementing appropriate tissue functionality tests to *normal-liver* and *diabetes-liver* and accounting for liver cell biomass changes, the two models were finalized and their characteristics are shown in **Figure 5.3**. Normal and disease liver models share a large number of reactions as well as the genes associated with the reactions. There are 112 reactions (4.68%) and 29 genes (2.08%) unique to the diabetic metabolic model. These 29 genes that were overexpressed only in the diabetic patients were further analyzed to reveal their potential association with diabetics. Using Metacore^{TM} [152], a bioinformatics online tool for pathway analysis, I analyzed how the input gene list intersects with Gene Ontology (GO) biological processes classification, and calculated the statistical significance of the intersection using a hypergeometric distribution [153]. The top three most enriched processes are: phosphagen and phosphocreatine metabolic process (*P-value* $5.00e^{-7}$), L-fucose metabolic process (*P-value* $4.16e^{-7}$) and long chain fatty acid metabolic process (*P-value* $8.32e^{-7}$). The phosphagens are energy storage compounds that can supply energy needs at a high rate. Phosphocreatine is one of the eight phosphagens and the most extensively studied phosphagen system [154]. They could synthesize ATP and decrease blood glucose content [154, 155], and having elevated levels of phosphagen-related genes may imply the body's effort to restore the high blood glucose level to normal. Similarly, impaired and dysregulation fatty acid metabolism has been reported for obesity, insulin resistance and T2DM [156, 157]. In diabetic cases, reduced insulin sensitivity in the liver promotes *de novo* fatty acid synthesis, and increased fats flow to the liver peripheral tissues.

Both effects cause excessive accumulation of hepatic fatty acids in the liver and eventually contribute to the development of fatty liver [157].

On the other hand, there are 37 reactions and 19 genes that showed evidence of presence only in normal liver cells, but not in diabetic liver cells. Metacore^{TM} was again used to characterize the genes. The top three most enriched GO processes for under-expressed genes are thiamine metabolic process (*P-value* $1.51e^{-4}$), pyrimidine-containing metabolic process (*P-value* $4.44e^{-3}$), and vitamin metabolic process (*P-value* $7.37e^{-3}$). Thiamine is a vitamin that plays a critical role in glucose metabolism [158]; it is known to be involved in the conversion of carbohydrates to glucose, and reduced thiamine metabolism may be due to the quick consumption and depletion of thiamine in diabetic patients as a result of high glucose metabolic activity.

Using Metacore^{TM}, I constructed a protein-protein interaction network using the gene products of the combined sets of over-expressed and under-expressed genes (**Figure 5.4**). The proteins either interact directly with each other (e.g., **KCRM** interacts with **NCX1**) or two proteins in the candidate lists interact with a common protein outside the candidate lists (e.g., both **NCX1** and **PCD2** interact with Carveolin2). The network consists of two sub-networks; a small one consists of two seeding genes and a larger one containing seven candidate genes. Both genes in the 2-gene network have been well known for their association with diabetes. The malic enzyme 1 (**ME1**) is a key regulator of fatty acid synthesis pathway and is highly susceptible to Type 2 diabetes [159]. It has been validated that **ME1** is a causal gene in for diabetic traits and genetically engineered mouse model with **ME1** knocked out were resistant to both diabetes and obesity development [160, 161]. Similarly, the mitochondrial branched-chain aminotransferase 2 (**BCAT2**) showed reduced expression in obesity and states of insulin resistance, which coincides with our finding that **BCAT2** was under- or not expressed in diabetic mice [162].

Genes in the larger network also have supporting literature proving their established relationship with diabetes; for instance, the sodium-calcium exchanger isoform 1 (**NCX1**) plays a key role in regulating cytoplasmic calcium required for insulin secretion [163], while vesicular monoamine transporter type 2 (**VMAT2**) is highly expressed in beta cells in the

pancreas and it has been used as a marker of sympathetic nerve terminals to quantify the amount of nerve loss from the islets of diabetic rats [164, 165].

5.3 CONCLUSIONS

In this chapter, I presented my work on the first liver model reconstruction for normal and diabetic mice. I started with improving and refining the state-of-art generic metabolic model for mice (Chapter 5.1), and then comparing and choosing the most suitable algorithm for reaction selection, eventually I successfully created relevant working liver models which possess the basic cell functions as well as liver-specific functions such as gluconeogenesis. This project ended with connecting relevant gene products to create a protein-protein interaction network. This protein-protein interaction network provides a valuable and comprehensive map to demonstrate connections among genes susceptible to diabetes, either being up- or downregulated. The genes and reactions unique to the *normal-liver* and to *diabetes-liver* demonstrated key metabolic differences between these two states, and provide a good starting point to better understand the metabolic pathways and mechanisms in diabetes, and to identify potential therapeutic targets for diabetic patients.

5.4 CHAPTER 5 FIGURES AND TABLES

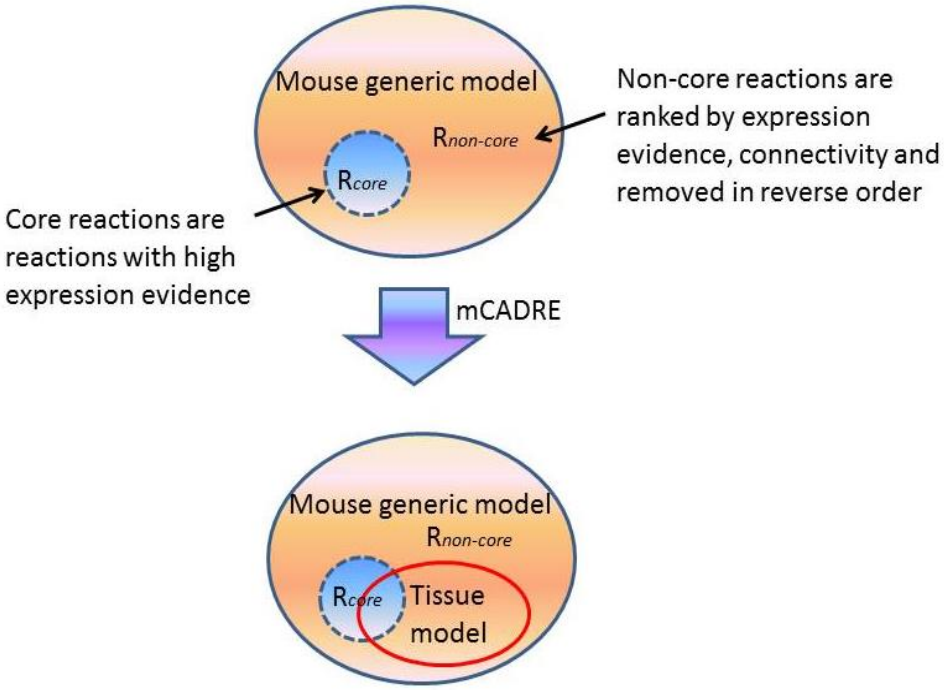


Figure 5.1 The tissue model building method (adapted from [128])

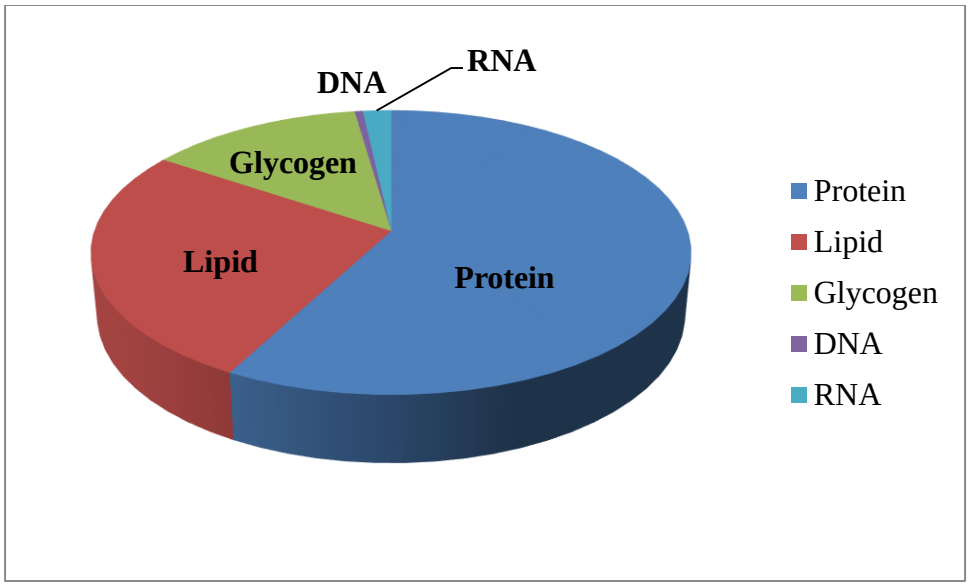


Figure 5.2 Hepatocyte cell composition

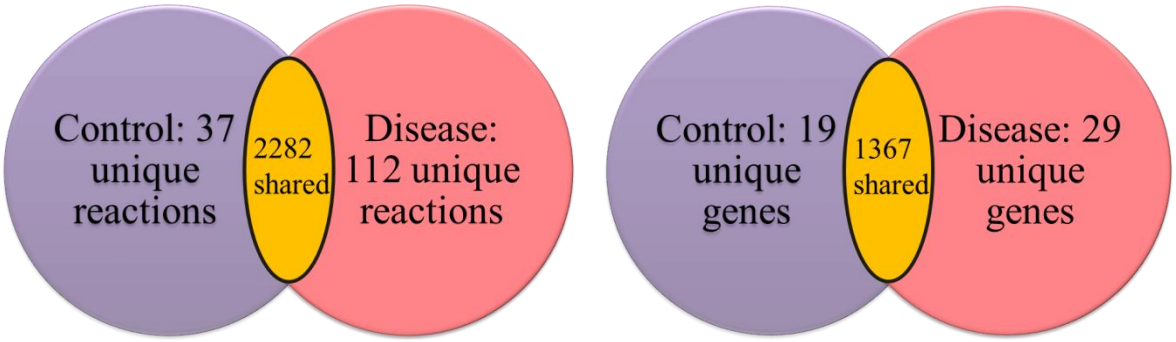


Figure 5.3 A) Reactions and B) genes in the final versions of normal and diabetes liver.

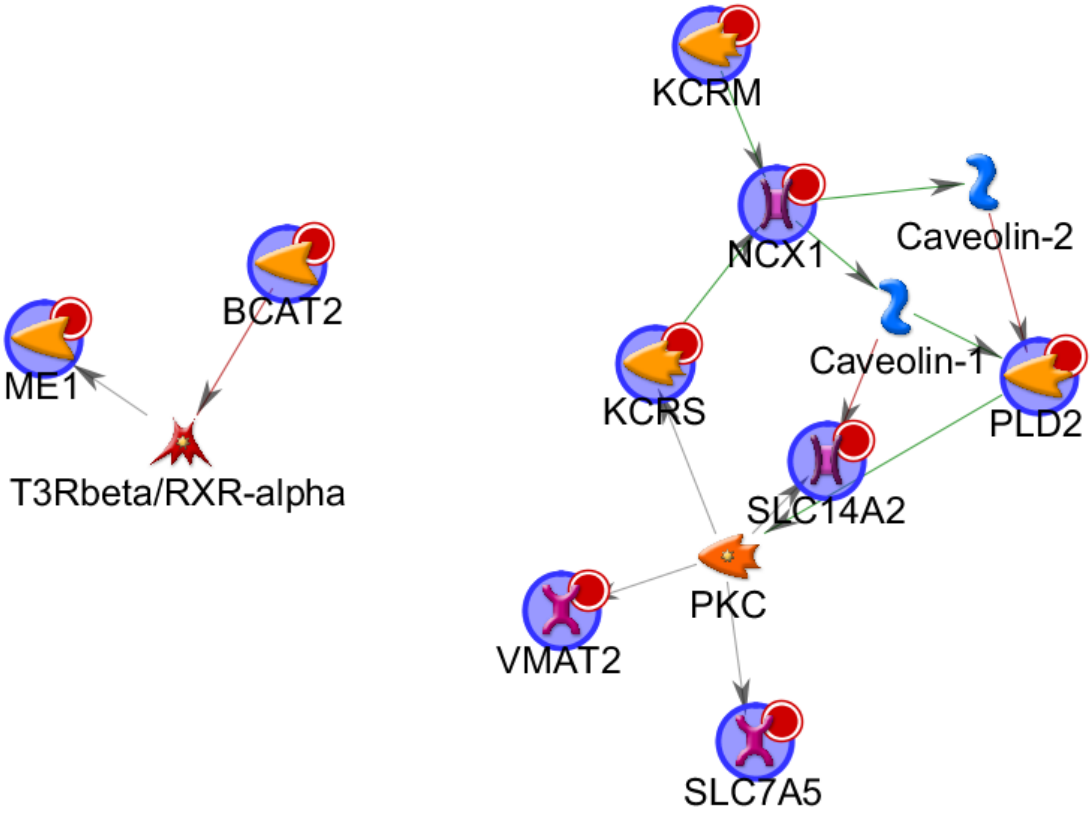


Figure 5.4 Interactions among diabetic-related genes

Name of metabolites	Number of metabolites	Category
Alanine, arginine, asparagine, aspartate, glutamine, glutamate, glycine, proline, serine	9	Non-essential amino acid
ATP, CTP, GTP, UTP, dATP, dCTP, dGTP, dTTP	8	Nucleotide
Ceramide, phosphatidylethanolamine, phosphatidylserine	3	Lipid

Table 5.1: list of metabolites current mouse model fails to produce

Authors of study (Year, GSE number)	GSM sample numbers	Number of samples in each study
Ackert-Bicknell *et al* (2006, GSE 5959) [166]	GSM138289 to GSM138291	3
Khetchoumian *et al* (2007, GSE9012) [167]	GSM228786 to GSM228790	5
Kozul *et al* (2008, GSE9630) [168]	GSM243352 to GSM243410	59
Tijet *et al* (2006, GSE 10082) [169]	GSM254871 to GSM254873 GSM254877 to GSM254883 254885	11
Hughes *et al* (2009,GSE11923)[170]	GSM301348 to GSM301395	48
MacLennan *et al* (2009, GSE12748) [171]	GSM319519, GSM319609 and GSM319903	3
Vollmers *et al* (2009, GSE13060) [172]	GSM327055 to GSM327129	24
Vollmers *et al* (2009, GSE13063) [172]	GSM327154 to GSM327161	8
Vollmers *et al* (2009, GSE13064) [172]	GSM327162 to GSM327169	8
Huang *et al* (2008, GSE13149) [173]	GSM329271 to GSM329295	25
Yates *et al* (2009, GSE15633) [174]	GSM391335 to GSM391340	6
Mohapatra *et al* (2010, GSE16207) [175]	GSM406976 to GSM406993	18
Tisserand et al (2011, GSE19675) [176]	GSM491305 to GSM491308, GSM491317 to GSM491321	9
Uehara *et al* (2011, GSE20562) [177]	GSM516651 to GSM491321	20
Lee *et al* (2010, GSE21224) [178]	GSM530635 to GSM530650	16
Lee *et al* (2010) N.A.	GSM541742 to GSM541757	16
Yu *et al* (2010, GSE21861) [179] GSM543655, 57, 59, 61	GSM543655, GSM543657, GSM543659, GSM543661	4
Dateki *et al* (2010, GSE22534) [180]	GSM559519, GSM559520	2
Dateki *et al* (2010, GSE22535) [180]	GSM559521 to GSM559524	4
Ding *et al* (2010, GSE 22879) [[181]	GSM565203 to GSM565206	4
Mongan *et al* (2010, GSE23780) [182]	GSM586822 to GSM586831	10
Duval et al (2010, GSE24031) [183]	GSM591473 to GSM591490	18
Edmonds et al (2011, GSE26695) [184]	GSM657144 to GSM657163	20
Pachikian et al (2011, GSE26986) [185]	GSM664751 to GSM664754	4
Zhang et al (2011, GSE27038) [186]	GSM665999 to GSM666001, GSM666005 to GSM666007	6

Table 5.2A Summary of microarray expression datasets included to reconstruct the control liver model

Authors of study (Year, GSE number)	Mouse strain	GSM sample numbers	Number of samples in each study
Stewart *et al* (2010, GSE24637) [187]	TALLYHO x C57BL6 F2	GSM607572 to GSM 607587	16
Davis *et al* (2011, GSE 30140) [188]	C57BL/6 x DBA/2 F2	GSM746336 to GSM 746551	264

Table 5.2B Summary of microarray expression datasets included to reconstruct the diabetic liver model

Input substrate reaction	Gluconeogenic substrate	Max glucose/substrate flux
EX_glyc(e)	Glycerol	0.484
EX_lac_L(e)	L-Lactate	0.343
EX_ala_L(e)	Alanine	0.336
EX_gln_L(e)	L-Glutamine	0.497
EX_pyr(e)	Pyruvate	0.272
EX_thr_L(e)	L-Threonine	0.420
EX_arg_L(e)	L-Arginine	0.584
EX_asn_L(e)	L-Asparagine	0.275
EX_asp_L(e)	L-Aspartate	0.319
EX_cys_L(e)	L-Cysteine	0.316
EX_glu_L(e)	L-Glutamate	0.527
EX_gly(e)	Glycine	0.017
EX_his_L(e)	L-Histidine	0.356
EX_ile_L(e)	L-Isoleucine	0.750
EX_met_L(e)	L-Methionine	0.303
EX_phe_L(e)	L-Phenylalanine	0.616
EX_pro_L(e)	L-Proline	0.620
EX_ser_L(e)	L-Serine	0.230
EX_trp_L(e)	L-Tryptophan	0.600
EX_tyr_L(e)	L-Tyrosine	0.685
EX_val_L(e)	L-Valine	0.500

Table 5.3 Gluconeogenic simulations. In all simulations, the optimized reaction is EX_glu-D(e).

Optimized reaction	Metabolite	Max ketone body/fatty acid flux
EX_acac(e)	Acetoacetate	1.50
EX_bhb(e)	B-hydroxybutyrate	1.48

Table 5.4A Ketogenic simulations

Optimized reaction	Tested sugar	mol ATP/mol sugar
ATPM	Fructose	22.8
ATPM	Galactose	22.8
ATPM	Mannose	22.8

Table 5.4B Alternative sugar simulations

Simulation	Optimized reaction	Reaction in the model	Metabolite uptake rate
Ammonia detoxification	Ammonia uptake	EX_nh4(e)	Maximized
Ethanol degradation	Ethanol uptake	EX_etoh(e)	Maximized
Amino acid degradation	Amino acid uptake	EX_asn_L(e)*	Maximized

Table 5.4C Ammonia, ethanol and amino acid degradations

* In the table L-Asparagine was used as a model amino acid to demonstrate the simulation process, all 20 amino acids were tested in both networks.

Optimized reaction	Reaction equation	Tested metabolite	Mol metabolite/mol glucose
DM_glycogen(c)	glycogen[c] →	Glycogen, cytosol	0.077
DM_chsterol(c)	chsterol[c] →	Cholesterol, cytosol	0.091

Table 5.4D Glycogen and cholesterol production simulations

Simulation	Reaction added	Reaction equation	Entrez Genes ID
Ketogenesis	EX_stcoa(e)	stcoa[e] <=>	N.A
Ketogenesis	STCOAt	stcoa[e] <=> stcoa[c]	N.A
Alternative sugar metabolism	FRUt4	na1[e] + fru[e] <=> na1[c] + fru[c]	230612.1
Alternative sugar metabolism	MAN6PI	man6p[c] <=> f6p[c]	110119.1
Gluconeogenesis	CSNAT2c	coa[c] + pcrn[c] <=> ppcoa[c] + crn[c]	12908.1
Glycogen production	Tyr_ggnt	Tyr-ggn[e] -> Tyr-ggn[c]	N.A

Table 5.5 Reactions added to the liver models after automatic pruning

Cell components	Component g/100g	Wet weight g/g	Dry weight g/g
Protein	17.3 [189]	0.172	0.519676
Lipids	7.95 [190]	0.079	0.23881
Glycogen	4 [191]	0.040	0.120156
Water	70.5 [192]	0.702	N.A.
DNA	0.15 [193]	0.001	0.004506
RNA	0.53 [194]	0.005	0.015921
Total	100.43	1.00	1.00

Table 5.6 Composition of the mouse liver cell

Amino acid	nmol/g tissue	Amino acid	nmol/g tissue
alanine	3.239	leucine	0.221
arginine	0.032	lysine	0.476
asparagine	0.143	methionine	0.062
aspartate	7.47	phenylalanine	0.087
cysteine	0.013	proline	0.194
glutamate	1.772	serine	0.957
glutamine	5.369	threonine	0.446
glycine	2.249	tryptophane	0.01
histidine	0.697	tyrosine	0.131
isoleucine	0.142	valine	0.215

Table 5.7A Amino acid of mouse liver cell. All values correspond to the rat data [195]

Lipids	%(g/g lipids)
Sphingomyelin	0.015
Cholesterol	0.052
Cholesterol esters	0.028
Monophosphoinosito	0.048
Phosphatidylethanolamine	0.093
Phosphatidylcholine and lisolecithin	0.246
Cardiolipin	0.017
Triacylglycerol	0.386
Total	0.885

Table 5.7B Lipid composition of the mouse liver tissue [196]

DNA	mol/mol DNA	RNA	mol/mol RNA
dAMP	0.3	AMP	0.18
dCMP	0.2	CMP	0.3
dGMP	0.2	GMP	0.34
dTMP	0.3	UMP	0.18

Table 5.7C Nucleotide composition of the mouse liver cell [197]

CHAPTER 6 CONCLUSIONS AND FUTURE DIRECTIONS

The last decade has witnessed an explosion in the amount of omics data generated by high-throughput technologies. The large amount of digital information enables a systems level understanding of the dependencies and correlations among molecular components. In my Ph.D. research work, I focused on utilizing a systems approach to analyze and characterize various diseases by exploring available transcriptomics, proteomics and metabolomics data.

In addition to analyzing a wide range of data types, I also studied a variety of diseases: human astrocytoma, mouse glioblastoma and mouse diabetes. In collecting and integrating proteomics and transcriptomics data from multiple lab sources, I utilized a uniform processing platform to increase sample-to-sample correlation and decrease heterogeneity across the data collected in different studies. To further mitigate biological noise and maximize disease effect, I also analyzed gene or protein expression levels in the context of biological network behaviors, which takes interactions among related gene or proteins into account and helps to link changes in gene expression to phenotype.

More specifically, I investigated and explored cancer aggressiveness and heterogeneity in the context of human astrocytoma, using transcriptomic data. Leveraging a large cohort of publicly available gene expression data sets, I have conducted the first meta-analysis that examines together the transcriptomes of three astrocytoma grades along with corresponding normal samples. I combined individual gene- and network-based approaches to identify meaningful patterns of expression within and between different grades. I quantified network dysregulation in each tumor grade and concluded that there is increasing inter-patient transcriptomic heterogeneity in more aggressive astrocytomas. Using a gene-based methodology, I also identified individual genes that exhibit monotonically increasing or decreasing expression with increased grade.

Having examined the heterogeneity in high grade astrocytoma, I moved on to develop pipelines to identify biomarkers indicative of clinical outcome. I developed an automated framework to screen all candidate networks in the pre-defined network database and discovered the erythropoietin network that is predictive of survival in HGA patients. This signature is known to

mediate neuroprotection through NF-κB signaling (EPONFκB). If the EPONFκB network is combined with previously reported signatures, the predictive power outperformed histology-based grading or those separated solely based on proneural/non-proneural status as predictors of survival. This pipeline is scalable and is capable of screening through a large number of networks rapidly and efficiently. For the EPONFκB network to move into clinical practice and adopted by medical doctors in regular diagnosis and prognosis, future experimental validation involving many more patients is required.

The mouse is usually the organism of choice to study human disease, and in the proteomics section of my dissertation, I analyzed large-scale proteomics data collected at different stages of mouse glioblastoma. Protein profiling was done on three genetically engineered mouse strains and over 2000 proteins were tracked at three tumor development stages: early, middle, and late. Individual proteins and networks that were most perturbed when GBM developed as a result of one or more major signaling pathways that exhibit frequent aberrant behavior in GBM become dysfunctional were identified. Ongoing experimental work is in progress to validate the selective protein targets of the genes.

Last, I developed the first genome-wide metabolic models for the C57BL/6J mouse liver, under normal and diabetic conditions, using mouse whole cell model as a starting point. I started with improving and refining the state-of-art generic metabolic model for mice, and then compared and chose the most suitable algorithm for reaction selection, eventually I successfully created relevant working liver models that possess the basic cell functions as well as liver-specific functions such as gluconeogenesis and amino acid degradation. Future work could follow the liver model reconstruction pipeline to create similar models for other tissues. For example, we could create adipose and muscle metabolic models for diabetic mice, and eventually add a connection compartment (i.e., the blood compartment) to form a multi-tissue model by integrating individual tissue models [131].

REFERENCES

1. Ideker, T., T. Galitski, and L. Hood, *A new approach to decoding life: systems biology.* Annu Rev Genomics Hum Genet, 2001. **2**: p. 343-72.
2. Tian, Q., N.D. Price, and L. Hood, *Systems cancer medicine: towards realization of predictive, preventive, personalized and participatory (P4) medicine.* J Intern Med, 2012. **271**(2): p. 111-21.
3. Ahn, A.C., et al., *The limits of reductionism in medicine: could systems biology offer an alternative?* PLoS Med, 2006. **3**(6): p. e208.
4. Auffray, C. and L. Hood, *Editorial: Systems biology and personalized medicine - the future is now.* Biotechnol J, 2012. **7**(8): p. 938-9.
5. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.
6. Ambros, V., *microRNAs: tiny regulators with great potential.* Cell, 2001. **107**(7): p. 823-6.
7. Black, D.L., *Mechanisms of alternative pre-messenger RNA splicing.* Annu Rev Biochem, 2003. **72**: p. 291-336.
8. Dhingra, V., et al., *New frontiers in proteomics research: a perspective.* Int J Pharm, 2005. **299**(1-2): p. 1-18.
9. Brower, V., *Proteomics: biology in the post-genomic era. Companies all over the world rush to lead the way in the new post-genomics race.* EMBO Rep, 2001. **2**(7): p. 558-60.
10. Domon, B. and R. Aebersold, *Mass spectrometry and protein analysis.* Science, 2006. **312**(5771): p. 212-7.
11. Wiese, S., et al., *Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research.* Proteomics, 2007. **7**(3): p. 340-50.
12. Samuelsson, L.M. and D.G. Larsson, *Contributions from metabolomics to fish research.* Mol Biosyst, 2008. **4**(10): p. 974-9.
13. Hwang, D., et al., *A systems approach to prion disease.* Mol Syst Biol, 2009. **5**: p. 252.
14. Oltvai, Z.N. and A.L. Barabasi, *Systems biology. Life's complexity pyramid.* Science, 2002. **298**(5594): p. 763-4.
15. http://www.mdanderson.org/patient-and-cancer-information/cancer-information/cancer-types/brain-cancer/index.html.
16. Chen, J., R.M. McKay, and L.F. Parada, *Malignant glioma: lessons from genomics, mouse models, and stem cells.* Cell, 2012. **149**(1): p. 36-47.
17. Chen, J., et al., *A restricted cell population propagates glioblastoma growth after chemotherapy.* Nature, 2012. **488**(7412): p. 522-6.
18. Louis, D.N., et al., *The 2007 WHO classification of tumours of the central nervous system.* Acta Neuropathol, 2007. **114**(2): p. 97-109.
19. Cheng, Y., et al., *Pilocytic astrocytomas do not show most of the genetic changes commonly seen in diffuse astrocytomas.* Histopathology, 2000. **37**(5): p. 437-444.
20. Kleihues, P. and H. Ohgaki, *Primary and secondary glioblastomas: from concept to clinical diagnosis.* Neuro Oncol, 1999. **1**(1): p. 44-51.
21. Shannon, P., et al., *Pathological and molecular progression of astrocytomas in a GFAP:12 V-Ha-Ras mouse astrocytoma model.* Am J Pathol, 2005. **167**(3): p. 859-67.

22. Nutt, C.L., et al., *Gene expression-based classification of malignant gliomas correlates better with survival than histological classification.* Cancer Res, 2003. **63**(7): p. 1602-7.
23. Shirahata, M., et al., *Gene expression-based molecular diagnostic system for malignant gliomas is superior to histological diagnosis.* Clinical Cancer Research, 2007. **13**(24): p. 7341-7356.
24. Freije, W.A., et al., *Gene expression profiling of gliomas strongly predicts survival.* Cancer Res, 2004. **64**(18): p. 6503-6510.
25. Phillips, H.S., et al., *Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis.* Cancer Cell, 2006. **9**(3): p. 157-173.
26. Wong, K.K., et al., *Expression analysis of juvenile pilocytic astrocytomas by oligonucleotide microarray reveals two potential subgroups.* Cancer Res, 2005. **65**(1): p. 76-84.
27. Sun, L.X., et al., *Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain.* Cancer Cell, 2006. **9**(4): p. 287-300.
28. Vitucci, M., D.N. Hayes, and C.R. Miller, *Gene expression profiling of gliomas: merging genomic and histopathological classification for personalised therapy.* Br J Cancer, 2011. **104**(4): p. 545-53.
29. Irizarry, R.A., et al., *Multiple-laboratory comparison of microarray platforms.* Nat Methods, 2005. **2**(5): p. 345-50.
30. Sung J, K.P., Ma S, Funk CC, Magis AT et al, *Multi-study Integration of Brain Cancer Transcriptomes Reveals Organ-Level Molecular Signatures.* PLoS Comput Biol, 2013. **9**(7): p. e1003148.
31. Grzmil, M. and B.A. Hemmings, *Deregulated signalling networks in human brain tumours.* Biochim Biophys Acta, 2010. **1804**(3): p. 476-83.
32. Eddy, J.A., et al., *Identifying Tightly Regulated and Variably Expressed Networks by Differential Rank Conservation (DIRAC).* Plos Computational Biology, 2010. **6**(5): p. -.
33. Leenstra, S., et al., *Molecular characterization of areas with low grade tumor or satellitosis in human malignant astrocytomas.* Cancer Res, 1992. **52**(6): p. 1568-72.
34. Alcantara Llaguno, S.R., J. Chen, and L.F. Parada, *Signaling in malignant astrocytomas: role of neural stem cells and its therapeutic implications.* Clinical Cancer Research, 2009. **15**(23): p. 7124-9.
35. Bonavia, R., et al., *Heterogeneity maintenance in glioblastoma: a social network.* Cancer Res, 2011. **71**(12): p. 4055-60.
36. Wypych, D. and P. Pomorski, *Calcium signaling in glioma cells--the role of nucleotide receptors.* Adv Exp Med Biol, 2013. **986**: p. 61-79.
37. Gabrilovich, D.I. and S. Nagaraj, *Myeloid-derived suppressor cells as regulators of the immune system.* Nat Rev Immunol, 2009. **9**(3): p. 162-74.
38. Fujita, M., et al., *COX-2 blockade suppresses gliomagenesis by inhibiting myeloid-derived suppressor cells.* Cancer Res, 2011. **71**(7): p. 2664-74.
39. Frey, H., et al., *Biological interplay between proteoglycans and their innate immune receptors in inflammation.* FEBS J, 2013.
40. Citri, A. and Y. Yarden, *EGF-ERBB signalling: towards the systems level.* Nat Rev Mol Cell Biol, 2006. **7**(7): p. 505-16.

41. Ekstrand, A.J., et al., *Genes for epidermal growth factor receptor, transforming growth factor alpha, and epidermal growth factor and their expression in human gliomas in vivo.* Cancer Res, 1991. **51**(8): p. 2164-72.
42. Stockhausen, M.T., K. Kristoffersen, and H.S. Poulsen, *The functional role of Notch signaling in human gliomas.* Neuro Oncol, 2010. **12**(2): p. 199-211.
43. Horikawa, I. and J.C. Barrett, *Transcriptional regulation of the telomerase hTERT gene as a target for cellular and viral oncogenic mechanisms.* Carcinogenesis, 2003. **24**(7): p. 1167-76.
44. Kyo, S. and M. Inoue, *Complex regulatory mechanisms of telomerase activity in normal and cancer cells: how can we apply them for cancer therapy?* Oncogene, 2002. **21**(4): p. 688-97.
45. Holand, K., F. Salm, and A. Arcaro, *The phosphoinositide 3-kinase signaling pathway as a therapeutic target in grade IV brain tumors.* Curr Cancer Drug Targets, 2011. **11**(8): p. 894-918.
46. Burke, P., K. Schooler, and H.S. Wiley, *Regulation of epidermal growth factor receptor signaling by endocytosis and intracellular trafficking.* Mol Biol Cell, 2001. **12**(6): p. 1897-910.
47. Desvergne, B., L. Michalik, and W. Wahli, *Transcriptional regulation of metabolism.* Physiol Rev, 2006. **86**(2): p. 465-514.
48. Menendez, J.A. and R. Lupu, *Fatty acid synthase and the lipogenic phenotype in cancer pathogenesis.* Nat Rev Cancer, 2007. **7**(10): p. 763-77.
49. Guo, D., et al., *An LXR agonist promotes glioblastoma cell death through inhibition of an EGFR/AKT/SREBP-1/LDLR-dependent pathway.* Cancer Discov, 2011. **1**(5): p. 442-56.
50. Zhu, J., et al., *Expression of R132H Mutational IDH1 in Human U87 Glioblastoma Cells Affects the SREBP1a Pathway and Induces Cellular Proliferation.* J Mol Neurosci, 2012.
51. Kim, W. and L.M. Liau, *IDH mutations in human glioma.* Neurosurg Clin N Am, 2012. **23**(3): p. 471-80.
52. Palty, R., et al., *SARAF inactivates the store operated calcium entry machinery to prevent excess calcium refilling.* Cell, 2012. **149**(2): p. 425-38.
53. Soboloff, J., et al., *STIM proteins: dynamic calcium signal transducers.* Nat Rev Mol Cell Biol, 2012. **13**(9): p. 549-65.
54. Castets, F., et al., *Zinedin, SG2NA, and striatin are calmodulin-binding, WD repeat proteins principally expressed in the brain.* J Biol Chem, 2000. **275**(26): p. 19970-7.
55. Mullins, F.M., et al., *STIM1 and calmodulin interact with Orai1 to induce Ca2+-dependent inactivation of CRAC channels.* Proc Natl Acad Sci U S A, 2009. **106**(36): p. 15495-500.
56. Giorgi, C., et al., *PML regulates apoptosis at endoplasmic reticulum by modulating calcium release.* Science, 2010. **330**(6008): p. 1247-51.
57. Peng, S.C., et al., *A novel role of CPEB3 in regulating EGFR gene transcription via association with Stat5b in neurons.* Nucleic Acids Res, 2010. **38**(21): p. 7446-57.
58. Hollander, M.C., G.M. Blumenthal, and P.A. Dennis, *PTEN loss in the continuum of common cancers, rare syndromes and mouse models.* Nat Rev Cancer, 2011. **11**(4): p. 289-301.

59. Roderick, H.L., J.D. Lechleiter, and P. Camacho, *Cytosolic phosphorylation of calnexin controls intracellular Ca(2+) oscillations via an interaction with SERCA2b.* J Cell Biol, 2000. **149**(6): p. 1235-48.
60. Lynes, E.M., et al., *Palmitoylated TMX and calnexin target to the mitochondria-associated membrane.* EMBO J, 2012. **31**(2): p. 457-70.
61. Rizzuto, R., et al., *Ca(2+) transfer from the ER to mitochondria: when, how and why.* Biochim Biophys Acta, 2009. **1787**(11): p. 1342-51.
62. Rizzuto, R., et al., *Mitochondria as sensors and regulators of calcium signalling.* Nat Rev Mol Cell Biol, 2012. **13**(9): p. 566-78.
63. Vander Heiden, M.G., L.C. Cantley, and C.B. Thompson, *Understanding the Warburg effect: the metabolic requirements of cell proliferation.* Science, 2009. **324**(5930): p. 1029-33.
64. Maloberti, P., et al., *Silencing the expression of mitochondrial acyl-CoA thioesterase I and acyl-CoA synthetase 4 inhibits hormone-induced steroidogenesis.* FEBS J, 2005. **272**(7): p. 1804-14.
65. Liu, Z., et al., *Drosophila Acyl-CoA synthetase long-chain family member 4 regulates axonal transport of synaptic vesicles and is required for synaptic development and transmission.* J Neurosci, 2011. **31**(6): p. 2052-63.
66. Phillips, H.S., et al., *Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis.* Cancer Cell, 2006. **9**(3): p. 157-73.
67. Markert, J.M., et al., *Differential gene expression profiling in human brain tumors.* Physiol Genomics, 2001. **5**(1): p. 21-33.
68. Ducker, C.E., et al., *Two N-myristoyltransferase isozymes play unique roles in protein myristoylation, proliferation, and apoptosis.* Mol Cancer Res, 2005. **3**(8): p. 463-76.
69. Ohgaki, H. and P. Kleihues, *Genetic pathways to primary and secondary glioblastoma.* Am J Pathol, 2007. **170**(5): p. 1445-53.
70. Godard, S., et al., *Classification of human astrocytic gliomas on the basis of gene expression: a correlated group of genes with angiogenic activity emerges as a strong predictor of subtypes.* Cancer Res, 2003. **63**(20): p. 6613-25.
71. Wu, Z.J., et al., *A model-based background adjustment for oligonucleotide expression arrays.* Journal of the American Statistical Association, 2004. **99**(468): p. 909-917.
72. Liu, W.M., et al., *Analysis of high density expression microarrays with signed-rank call algorithms.* Bioinformatics, 2002. **18**(12): p. 1593-1599.
73. 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. 15545-15550.
74. Byrne, T.N., *Cognitive sequelae of brain tumor treatment.* Curr Opin Neurol, 2005. **18**(6): p. 662-6.
75. Celik, M., et al., *Erythropoietin prevents motor neuron apoptosis and neurologic disability in experimental spinal cord ischemic injury.* Proc Natl Acad Sci U S A, 2002. **99**(4): p. 2258-63.
76. Keswani, S.C., et al., *A novel endogenous erythropoietin mediated pathway prevents axonal degeneration.* Ann Neurol, 2004. **56**(6): p. 815-26.

77. Shingo, T., et al., *Erythropoietin regulates the in vitro and in vivo production of neuronal progenitors by mammalian forebrain neural stem cells.* J Neurosci, 2001. **21**(24): p. 9733-43.
78. Erbayraktar, S., et al., *Carbamylated erythropoietin reduces radiosurgically-induced brain injury.* Mol Med, 2006. **12**(4-6): p. 74-80.
79. Pinel, S., et al., *Erythropoietin-induced reduction of hypoxia before and during fractionated irradiation contributes to improvement of radioresponse in human glioma xenografts.* Int J Radiat Oncol Biol Phys, 2004. **59**(1): p. 250-9.
80. Nico, B., et al., *Epo is involved in angiogenesis in human glioma.* J Neurooncol, 2011. **102**(1): p. 51-8.
81. Hassouna, I., et al., *Erythropoietin augments survival of glioma cells after radiation and temozolomide.* Int J Radiat Oncol Biol Phys, 2008. **72**(3): p. 927-34.
82. Li, M., et al., *Widespread RNA and DNA sequence differences in the human transcriptome.* Science, 2011. **333**(6038): p. 53-8.
83. Institute, B. *iTRAQ*. Available from: http://www.broadinstitute.org/scientific-community/science/platforms/proteomics/itraq.
84. Picotti, P. and R. Aebersold, *Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions.* Nat Methods, 2012. **9**(6): p. 555-66.
85. Gillet, L.C., et al., *Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis.* Mol Cell Proteomics, 2012. **11**(6): p. O111 016717.
86. Tu, Y., G. Stolovitzky, and U. Klein, *Quantitative noise analysis for gene expression microarray experiments.* Proc Natl Acad Sci U S A, 2002. **99**(22): p. 14031-6.
87. Brennan, C., *Genomic profiles of glioma.* Curr Neurol Neurosci Rep, 2011. **11**(3): p. 291-7.
88. Backlund, L.M., et al., *Short postoperative survival for glioblastoma patients with a dysfunctional Rb1 pathway in combination with no wild-type PTEN.* Clin Cancer Res, 2003. **9**(11): p. 4151-8.
89. Kranenburg, O., *The KRAS oncogene: past, present, and future.* Biochim Biophys Acta, 2005. **1756**(2): p. 81-2.
90. Wang, S.I., et al., *Somatic mutations of PTEN in glioblastoma multiforme.* Cancer Res, 1997. **57**(19): p. 4183-6.
91. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.
92. Wilcoxon, F., *Individual comparisons of grouped data by ranking methods.* J Econ Entomol, 1946. **39**: p. 269.
93. Raju, T.N., *William Sealy Gosset and William A. Silverman: two "students" of science.* Pediatrics, 2005. **116**(3): p. 732-5.
94. Bult, C.J., et al., *The Mouse Genome Database (MGD): integrating biology with the genome.* Nucleic Acids Res, 2004. **32**(Database issue): p. D476-81.
95. Eppig, J.T., et al., *The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse.* Nucleic Acids Res, 2012. **40**(Database issue): p. D881-6.

96. Yang, Z., T. Zhao, and Y. Liu, *Upregulation of tumor suppressor WWOX promotes immune response in glioma.* Cell Immunol, 2013. **285**(1-2): p. 1-5.
97. Saied, I.T. and A.M. Shamsuddin, *Up-regulation of the tumor suppressor gene p53 and WAF1 gene expression by IP6 in HT-29 human colon carcinoma cell line.* Anticancer Res, 1998. **18**(3A): p. 1479-84.
98. Zindy, P.J., et al., *Upregulation of the tumor suppressor gene menin in hepatocellular carcinomas and its significance in fibrogenesis.* Hepatology, 2006. **44**(5): p. 1296-307.
99. Rak, J., et al., *Mutant ras oncogenes upregulate VEGF/VPF expression: implications for induction and inhibition of tumor angiogenesis.* Cancer Res, 1995. **55**(20): p. 4575-80.
100. Okada, F., et al., *Impact of oncogenes in tumor angiogenesis: mutant K-ras up-regulation of vascular endothelial growth factor/vascular permeability factor is necessary, but not sufficient for tumorigenicity of human colorectal carcinoma cells.* Proc Natl Acad Sci U S A, 1998. **95**(7): p. 3609-14.
101. Kuo, W.P., et al., *Gene expression levels in different stages of progression in oral squamous cell carcinoma.* Proc AMIA Symp, 2002: p. 415-9.
102. Noch, E. and K. Khalili, *Molecular mechanisms of necrosis in glioblastoma: the role of glutamate excitotoxicity.* Cancer Biol Ther, 2009. **8**(19): p. 1791-7.
103. Raza, S.M., et al., *Necrosis and glioblastoma: a friend or a foe? A review and a hypothesis.* Neurosurgery, 2002. **51**(1): p. 2-12; discussion 12-3.
104. Song, Y., et al., *Evolutionary etiology of high-grade astrocytomas.* Proc Natl Acad Sci U S A, 2013. **110**(44): p. 17933-8.
105. Matlashewski, G., et al., *Isolation and characterization of a human p53 cDNA clone: expression of the human p53 gene.* EMBO J, 1984. **3**(13): p. 3257-62.
106. Rasheed, B.K., et al., *Alterations of the TP53 gene in human gliomas.* Cancer Res, 1994. **54**(5): p. 1324-30.
107. Naccarati, A., et al., *Mutations and polymorphisms in TP53 gene--an overview on the role in colorectal cancer.* Mutagenesis, 2012. **27**(2): p. 211-8.
108. Kato, H., et al., *Functional evaluation of p53 and PTEN gene mutations in gliomas.* Clin Cancer Res, 2000. **6**(10): p. 3937-43.
109. Knobbe, C.B., A. Merlo, and G. Reifenberger, *Pten signaling in gliomas.* Neuro Oncol, 2002. **4**(3): p. 196-211.
110. Tohma, Y., et al., *PTEN (MMAC1) mutations are frequent in primary glioblastomas (de novo) but not in secondary glioblastomas.* J Neuropathol Exp Neurol, 1998. **57**(7): p. 684-9.
111. Munoz, J., et al., *Homozygous deletion and expression of PTEN and DMBT1 in human primary neuroblastoma and cell lines.* Int J Cancer, 2004. **109**(5): p. 673-9.
112. Bostrom, J., et al., *Mutation of the PTEN (MMAC1) tumor suppressor gene in a subset of glioblastomas but not in meningiomas with loss of chromosome arm 10q.* Cancer Res, 1998. **58**(1): p. 29-33.
113. Schellenberger, J., et al., *Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0.* Nat Protoc, 2011. **6**(9): p. 1290-307.
114. Edwards, J.S. and B.O. Palsson, *Systems properties of the Haemophilus influenzae Rd metabolic genotype.* J Biol Chem, 1999. **274**(25): p. 17410-6.

115. Feist, A.M., et al., *Reconstruction of biochemical networks in microorganisms.* Nat Rev Microbiol, 2009. **7**(2): p. 129-43.
116. Thiele, I. and B.O. Palsson, *A protocol for generating a high-quality genome-scale metabolic reconstruction.* Nat Protoc, 2010. **5**(1): p. 93-121.
117. Hood, L., et al., *Systems biology and new technologies enable predictive and preventative medicine.* Science, 2004. **306**(5696): p. 640-3.
118. Panchal, S.K. and L. Brown, *Rodent models for metabolic syndrome research.* J Biomed Biotechnol, 2011. **2011**: p. 351982.
119. *Summaries for patients. The effect of diet and exercise or metformin on the metabolic syndrome.* Ann Intern Med, 2005. **142**(8): p. I46.
120. Sigurdsson, M.I., et al., *A detailed genome-wide reconstruction of mouse metabolism based on human Recon 1.* BMC Syst Biol, 2010. **4**: p. 140.
121. Duarte, N.C., et al., *Global reconstruction of the human metabolic network based on genomic and bibliomic data.* Proc Natl Acad Sci U S A, 2007. **104**(6): p. 1777-82.
122. Thiele, I., et al., *A community-driven global reconstruction of human metabolism.* Nat Biotechnol, 2013. **31**(5): p. 419-25.
123. Hascup, K.N., et al., *Differential levels of glutamate dehydrogenase 1 (GLUD1) in Balb/c and C57BL/6 mice and the effects of overexpression of the Glud1 gene on glutamate release in striatum.* ASN Neuro, 2011. **3**(2).
124. Vetterli, L., et al., *Delineation of glutamate pathways and secretory responses in pancreatic islets with beta-cell-specific abrogation of the glutamate dehydrogenase.* Mol Biol Cell, 2012. **23**(19): p. 3851-62.
125. Patete, P., et al., *A multi-tissue mass-spring model for computer assisted breast surgery.* Med Eng Phys, 2012.
126. Lewis, N.E., et al., *A dynamic cpSRP43-Albino3 interaction mediates translocase regulation of chloroplast signal recognition particle (cpSRP)-targeting components.* J Biol Chem, 2010. **285**(44): p. 34220-30.
127. Baek, K.J., et al., *Phospholipase Cdelta1 is a guanine nucleotide exchanging factor for transglutaminase II (Galpha h) and promotes alpha 1B-adrenoreceptor-mediated GTP binding and intracellular calcium release.* J Biol Chem, 2001. **276**(8): p. 5591-7.
128. Jerby, L., T. Shlomi, and E. Ruppin, *Computational reconstruction of tissue-specific metabolic models: application to human liver metabolism.* Mol Syst Biol, 2010. **6**: p. 401.
129. Alvarez, I., et al., *Progress of National Multi-tissue Bank in Uruguay in the International Atomic Energy Agency (IAEA) Tissue Banking Programme.* Cell Tissue Bank, 2003. **4**(2-4): p. 173-8.
130. Uchida, N., F.Y. Leung, and C.J. Eaves, *Liver and marrow of adult mdr-1a/1b(-/-) mice show normal generation, function, and multi-tissue trafficking of primitive hematopoietic cells.* Exp Hematol, 2002. **30**(8): p. 862-9.
131. Bordbar, A., et al., *A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology.* BMC Syst Biol, 2011. **5**: p. 180.
132. ; Available from: http://www.umm.edu/altmed/articles/diabetes-000049.htm.
133. Reaven, G.M., *Why Syndrome X? From Harold Himsworth to the insulin resistance syndrome.* Cell Metab, 2005. **1**(1): p. 9-14.

134. Rich, S.S. and R.N. Bergman, *The genetic basis of glucose homeostasis.* Curr Diabetes Rev, 2005. **1**(3): p. 221-6.
135. Lakka, H.M., et al., *The metabolic syndrome and total and cardiovascular disease mortality in middle-aged men.* JAMA, 2002. **288**(21): p. 2709-16.
136. Esposito, K., A. Ceriello, and D. Giugliano, *Diet and the metabolic syndrome.* Metab Syndr Relat Disord, 2007. **5**(4): p. 291-6.
137. Blake, J.A., et al., *The Mouse Genome Database genotypes::phenotypes.* Nucleic Acids Res, 2009. **37**(Database issue): p. D712-9.
138. Bryant, C.D., et al., *Behavioral Differences among C57BL/6 Substrains: Implications for Transgenic and Knockout Studies.* Journal of Neurogenetics, 2008. **22**(4): p. 315-331.
139. Zurita, E., et al., *Genetic polymorphisms among C57BL/6 mouse inbred strains.* Transgenic Research, 2011. **20**(3): p. 481-489.
140. Wang, Y., J.A. Eddy, and N.D. Price, *Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE.* BMC Syst Biol, 2012. **6**: p. 153.
141. Affymetrix.com. *Mouse430_2 Probe Sequences, FASTA* 2008; Available from: http://www.affymetrix.com/estore/support/file_download.affx?onloadforward=/analysis/downloads/data/Mouse430_2.probe_fasta.zip&_requestid=173471.
142. Affymetrix.com, *Mouse Genome 430 2.0 Array.*
143. Biomart.com. *ID converter*. Available from: http://central.biomart.org/converter/#!/ID_converter/gene_ensembl_config_2.
144. Carlsson, P. and L. Kjellen, *Heparin biosynthesis.* Handb Exp Pharmacol, 2012(207): p. 23-41.
145. Gerich, J.E., et al., *Renal gluconeogenesis: its importance in human glucose homeostasis.* Diabetes Care, 2001. **24**(2): p. 382-91.
146. Stipanuk, M.H., *Biochemical and physiological aspects of human nutrition*2000, Philadelphia: W.B. Saunders. xxx, 1007 p.
147. Heinz, F., W. Lamprecht, and J. Kirsch, *Enzymes of fructose metabolism in human liver.* J Clin Invest, 1968. **47**(8): p. 1826-32.
148. Jeppesen, J.B., et al., *Lactate metabolism in chronic liver disease.* Scand J Clin Lab Invest, 2013.
149. Herman, R.H., *Mannose metabolism. I.* Am J Clin Nutr, 1971. **24**(4): p. 488-98.
150. Rennie, M.J. and K.D. Tipton, *Protein and amino acid metabolism during and after exercise and the effects of nutrition.* Annu Rev Nutr, 2000. **20**: p. 457-83.
151. Selvarasu, S., et al., *Genome-scale modeling and in silico analysis of mouse cell metabolic network.* Mol Biosyst, 2010. **6**(1): p. 152-61.
152. Reuters, T. *Metacore*. Available from: http://thomsonreuters.com/metacore/.
153. Metacore. *Evaluating statistical significance of pathways and network in MetaCore*. Available from: https://portal.genego.com/help/P-value_calculations.pdf.
154. Ellington, W.R., *Evolution and physiological roles of phosphagen systems.* Annu Rev Physiol, 2001. **63**: p. 289-325.
155. Examiner.com. *Type 1 diabetes and CrossFit*. 2012; Available from: http://www.examiner.com/article/type-1-diabetes-and-crossfit.
156. Blaak, E.E., *Fatty acid metabolism in obesity and type 2 diabetes mellitus.* Proc Nutr Soc, 2003. **62**(3): p. 753-60.

157. Postic, C. and J. Girard, *Contribution of de novo fatty acid synthesis to hepatic steatosis and insulin resistance: lessons from genetically engineered mice.* J Clin Invest, 2008. **118**(3): p. 829-38.
158. Lonsdale, D., *A review of the biochemistry, metabolism and clinical benefits of thiamin(e) and its derivatives.* Evid Based Complement Alternat Med, 2006. **3**(1): p. 49-59.
159. Zhou, S.L., et al., *Differential expression analysis of porcine MDH1, MDH2 and ME1 genes in adipose tissues.* Genet Mol Res, 2012. **11**(2): p. 1254-9.
160. Yang, X., et al., *Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks.* Nat Genet, 2009. **41**(4): p. 415-23.
161. Zhong, H., et al., *Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes.* PLoS Genet, 2010. **6**(5): p. e1000932.
162. Melnik, B.C., *Leucine signaling in the pathogenesis of type 2 diabetes and obesity.* World J Diabetes, 2012. **3**(3): p. 38-53.
163. Hamming, K.S., et al., *Inhibition of beta-cell sodium-calcium exchange enhances glucose-dependent elevations in cytoplasmic calcium and insulin secretion.* Diabetes, 2010. **59**(7): p. 1686-93.
164. Mei, Q., et al., *Early, selective, and marked loss of sympathetic nerves from the islets of BioBreeder diabetic rats.* Diabetes, 2002. **51**(10): p. 2997-3002.
165. Freeby, M., et al., *VMAT2 quantitation by PET as a biomarker for beta-cell mass in health and disease.* Diabetes Obes Metab, 2008. **10 Suppl 4**: p. 98-108.
166. Ackert-Bicknell, C.L., et al., *A chromosomal inversion within a quantitative trait locus has a major effect on adipogenesis and osteoblastogenesis.* Ann N Y Acad Sci, 2007. **1116**: p. 291-305.
167. Khetchoumian, K., et al., *Loss of Trim24 (Tif1alpha) gene function confers oncogenic activity to retinoic acid receptor alpha.* Nat Genet, 2007. **39**(12): p. 1500-6.
168. Kozul, C.D., et al., *Laboratory diet profoundly alters gene expression and confounds genomic analysis in mouse liver and lung.* Chem Biol Interact, 2008. **173**(2): p. 129-40.
169. Tijet, N., et al., *Aryl hydrocarbon receptor regulates distinct dioxin-dependent and dioxin-independent gene batteries.* Mol Pharmacol, 2006. **69**(1): p. 140-53.
170. Hughes, M.E., et al., *Harmonics of circadian gene transcription in mammals.* PLoS Genet, 2009. **5**(4): p. e1000442.
171. MacLennan, N.K., et al., *Weighted gene co-expression network analysis identifies biomarkers in glycerol kinase deficient mice.* Mol Genet Metab, 2009. **98**(1-2): p. 203-14.
172. Vollmers, C., et al., *Time of feeding and the intrinsic circadian clock drive rhythms in hepatic gene expression.* Proc Natl Acad Sci U S A, 2009. **106**(50): p. 21453-8.
173. Li, T., et al., *Multi-stage analysis of gene expression and transcription regulation in C57/B6 mouse liver development.* Genomics, 2009. **93**(3): p. 235-42.
174. Yates, M.S., et al., *Genetic versus chemoprotective activation of Nrf2 signaling: overlapping yet distinct gene expression profiles between Keap1 knockout and triterpenoid-treated mice.* Carcinogenesis, 2009. **30**(6): p. 1024-31.
175. Mohapatra, S.K., et al., *Modulation of hepatic PPAR expression during Ft LVS LPS-induced protection from Francisella tularensis LVS infection.* BMC Infect Dis, 2010. **10**: p. 10.

176. Tisserand, J., et al., *Tripartite motif 24 (Trim24/Tif1alpha) tumor suppressor protein is a novel negative regulator of interferon (IFN)/signal transducers and activators of transcription (STAT) signaling pathway acting through retinoic acid receptor alpha (Raralpha) inhibition.* J Biol Chem, 2011. **286**(38): p. 33369-79.
177. Uehara, Y., et al., *Gene expression profiles in mouse liver after long-term low-dose-rate irradiation with gamma rays.* Radiat Res, 2010. **174**(5): p. 611-7.
178. Lee, J.S., et al., *Transcriptional ontogeny of the developing liver.* BMC Genomics, 2012. **13**: p. 33.
179. Yu, J.H., et al., *The transcription factors signal transducer and activator of transcription 5A (STAT5A) and STAT5B negatively regulate cell proliferation through the activation of cyclin-dependent kinase inhibitor 2b (Cdkn2b) and Cdkn1a expression.* Hepatology, 2010. **52**(5): p. 1808-18.
180. Dateki, M., et al., *Adaptive gene regulation of pyruvate dehydrogenase kinase isoenzyme 4 in hepatotoxic chemical-induced liver injury and its stimulatory potential for DNA repair and cell proliferation.* J Recept Signal Transduct Res, 2011. **31**(1): p. 85-95.
181. Ding, B.S., et al., *Inductive angiocrine signals from sinusoidal endothelium are required for liver regeneration.* Nature, 2010. **468**(7321): p. 310-5.
182. Mongan, M.A., et al., *A novel statistical algorithm for gene expression analysis helps differentiate pregnane X receptor-dependent and independent mechanisms of toxicity.* PLoS One, 2010. **5**(12): p. e15595.
183. Duval, C., et al., *Adipose tissue dysfunction signals progression of hepatic steatosis towards nonalcoholic steatohepatitis in C57BL/6 mice.* Diabetes, 2010. **59**(12): p. 3181-91.
184. Edmonds, R.D., et al., *Transcriptomic response of murine liver to severe injury and hemorrhagic shock: a dual-platform microarray analysis.* Physiol Genomics, 2011. **43**(20): p. 1170-83.
185. Pachikian, B.D., et al., *Hepatic n-3 polyunsaturated fatty acid depletion promotes steatosis and insulin resistance in mice: genomic analysis of cellular targets.* PLoS One, 2011. **6**(8): p. e23365.
186. Zhang, K., et al., *The unfolded protein response transducer IRE1alpha prevents ER stress-induced hepatic steatosis.* EMBO J, 2011. **30**(7): p. 1357-75.
187. Stewart, T.P., et al., *Genetic and genomic analysis of hyperlipidemia, obesity and diabetes using (C57BL/6J x TALLYHO/JngJ) F2 mice.* BMC Genomics, 2010. **11**: p. 713.
188. Davis, R.C., et al., *Systems genetics of susceptibility to obesity-induced diabetes in mice.* Physiol Genomics, 2012. **44**(1): p. 1-13.
189. Bernard, A., et al., *Hyperinsulinemia induced by canine distemper virus infection of mice and its correlation with the appearance of obesity.* Comp Biochem Physiol B, 1988. **91**(4): p. 691-6.
190. Kelley, D.S., et al., *Fatty acid composition of liver, adipose tissue, spleen, and heart of mice fed diets containing t10, c12-, and c9, t11-conjugated linoleic acid.* Prostaglandins Leukot Essent Fatty Acids, 2006. **74**(5): p. 331-8.
191. Hubner, R.H., et al., *Dysfunctional glycogen storage in a mouse model of alpha1-antitrypsin deficiency.* Am J Respir Cell Mol Biol, 2009. **40**(2): p. 239-47.

192. Reinoso, R.F., B.A. Telfer, and M. Rowland, *Tissue water content in rats measured by desiccation.* J Pharmacol Toxicol Methods, 1997. **38**(2): p. 87-92.
193. Promega.com. *Purifying RNA and mRNA*. Available from: http://www.promega.com/~/media/files/resources/product%20guides/rna%20analysis%20notebook/purifyingrna.pdf?la=en.
194. Jones, L.D., M.K. Nielsen, and R.A. Britton, *Genetic variation in liver mass, body mass, and liver:body mass in mice.* J Anim Sci, 1992. **70**(10): p. 2999-3006.
195. Ana Triguero, T.B., Concha GarcÍa, Inmaculada R. Puertes, Juan Sastre, Juan R. ViÑa, *Liver intracellular L-cysteine concentration is maintained after inhibition of the trans-sulfuration pathway by propargylglycine in rats.* British Journal of Nutrition, 2007. **78**(05): p. 823-831.
196. Nelson, G.J., *The lipid composition of normal mouse liver.* Journal of lipid research, 1962. **3**(1): p. 256-262.
197. Sheikh, K., J. Forster, and L.K. Nielsen, *Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of Mus musculus.* Biotechnol Prog, 2005. **21**(1): p. 112-21.