ANALYSIS OF TRANSCRIPTOME DATA FOR THE EVALUATION OF METABOLIC DEREGULATION IN CANCER

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ILIR SHERAJ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
BIOLOGY

JUNE 2021

Approval of the thesis:

ANALYSIS OF TRANSCRIPTOME DATA FOR THE EVALUATION OF METABOLIC DEREGULATION IN CANCER

submitted by **ILIR SHERAJ** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy** i**n Biology, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr Ayşe Gül Gözen
Head of the Department, **Biology** _____

Prof. Dr. Sreeparna Banerjee
Supervisor, **Biology, METU** _____

Examining Committee Members:

Prof. Dr. Nülüfer Tülin Güray
Biology, METU _____

Prof. Dr. Sreeparna Banerjee
Biology, METU _____

Assoc. Prof. Dr. Özlen Konu Karakayalı
Molecular Biology and Genetics, Bilkent Uni. _____

Assoc. Prof. Dr. Nurcan Tunçbağ
Chemical and Biological Engineering, Koç Uni _____

Assist. Prof. Dr. Ahmet Acar
Biology, METU _____

Date: 28.06.2021

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name Last name : Ilir Sheraj

Signature :

ABSTRACT

ANALYSIS OF TRANSCRIPTOME DATA FOR THE EVALUATION OF METABOLIC DEREGULATION IN CANCER

Sheraj, Ilir
Doctor of Philosophy, Biology
Supervisor : Prof. Dr. Sreeparna Banerjee

June 2021, 180 pages

The great potential of cancer metabolism in tumor development as well as treatment has been appreciated in recent years. In this study, an extensive list of enzymes relevant to central cellular metabolism and metabolite transporters were manually curated. A pan cancer differential expression of these genes between tumor and their normal counterpart was analyzed in an attempt to better understand tumor bioenergetics, catabolism and anabolism in general. Major deregulation of these genes, particularly the enzymes involved in NADPH metabolism was observed in Liver Hepatocellular Carcinoma (LIHC). LIHC is an extremely aggressive treatment resistant disease with poor prognosis and overall survival (OS), probably due to the ability of the organ to detoxify chemotherapy drugs. AKR1B10 is an NADPH utilizing enzyme whose expression was increased from a very early stage in most LIHC patients and was associated with low OS. However, mechanistic underpinnings of this association are unknown. To address this and the context in which it occurs we hypothesized a potential connection of AKR1B10 with the pentose phosphate pathway (PPP), a glucose metabolizing pathway that generates NADPH. Using various bioinformatics and computational approaches, we found that AKR1B10 works in concert with PPP and a number of other detoxifying enzymes to

enhance tumor aggressiveness. In addition, we report the use of latest tools in machine learning to build multi-gene signature prognostic models for highly significant LIHC patient stratification.

Keywords: Tumor Metabolism, LIHC, AKR1B10, TCGA

ÖZ

KANSERDE METABOLİK DEREGÜLASYONUN DEĞERLENDİRİLMESİ İÇİN TRANSKRİPTOM VERİLERİNİN ANALİZİ

Sheraj, Ilir
Doktora, Biyoloji
Tez Yöneticisi: Prof. Dr. Sreeparna Banerjee

Haziran 2018, 180 sayfa

Kanser metabolizmasının tümör gelişiminde olduğu kadar tedavideki büyük potansiyeli de son yıllarda takdir edilmiştir. Bu çalışmada, merkezi hücresel metabolizma ve metabolit taşıyıcıları ile ilgili kapsamlı bir enzim listesi manuel olarak küratörlüğünü yaptı. Bu genlerin tümör ve normal muadilleri arasında bir pan kanser diferansiyel ekspresyonu, tümör biyoenerjisini, katabolizmasını ve genel olarak anabolizmi daha iyi anlamak amacıyla analiz edildi. Bu genlerin, özellikle NADPH metabolizmasında yer alan enzimlerin, Karaciğer Hepatoselüler Karsinomda (LIHC) büyük ölçüde düzensizleşmesi gözlendi. LIHC, muhtemelen organın kemoterapi ilaçlarını detoksifiye etme yeteneğinden dolayı, kötü prognoz ve genel sağkalıma sahip, son derece agresif, tedaviye dirençli bir hastalıktır. AKR1B10, çoğu LIHC hastasında ekspresyonu çok erken bir aşamada artan ve düşük OS ile ilişkili olan, NADPH kullanan bir enzimdir. Bununla birlikte, bu ilişkinin mekanik temelleri bilinmemektedir. Bunu ve meydana geldiği bağlamı ele almak için, NADPH üreten bir glikoz metabolize edici yol olan pentoz fosfat yolu (PPP) ile AKR1B10un potansiyel bir bağlantısını varsaydık. Çeşitli biyoinformatik ve hesaplama yaklaşımlarını kullanarak, AKR1B10un PPP ve bir dizi başka

detoksifiye edici enzimle uyumlu çalıştığını ve tümörün agresifliğini arttırdığını bulduk. Ek olarak, son derece önemli LIHC hasta sınıflandırması için çok gen imzalı prognostik modeller oluşturmak için makine öğreniminde en son araçların kullanımını rapor ediyoruz.

Anahtar Kelimeler: Tümör Metabolizması, LIHC, AKR1B10, TCGA

To all those unlucky souls represented by 0’s and 1’s in KM plots, and METTMFs

ACKNOWLEDGMENTS

I would like to express my deep gratitude to my supervisor, Prof Sreeparna Banerjee for her guidance, advice, criticism and especially the freedom to do often times whatever I wished and pulling me back into tracks whenever I would go too far, which happens quite often with me. I would also like to thank the CST members, Prof Tulin Guray and Prof Ozlen Konu for their valuable and constructive advice during CSTs. Those meetings always made a difference! I would also like to thank my family for their support and patience since my busy schedule allowed me to visit them in person only once in the last 6 years. I would also like to thank members of Banerjee Lab.

This study is all based on bioinformatics and computation, a field that I started as a complete novice nearly 3 and a half years ago. Therefore, it would have been impossible without the support of hundreds or thousands of people who shared their codes and answered questions in Stack Overflow and Bioconductor.

Last but not least, special thanks go to many friends who have accompanied me for so many years in this city, in particular engineers, physicists and mathematicians that increased my curiosity for the amazing field of quantitative sciences and totally changed my view of Biology. A few special names include Sumaya Asif, Mohammad Ragai, Saed Eusufzae, Majid Akbar, Jawad Nizami, Bakht Afridi, Tanvir Shah, Indrit Nallbani, Chaner Abdullah, and many others. Countless hours sipping coffee and discussing almost everything from Science, Technology, Mathematics, Physics, Middle Eastern Politics, and of course dollar exchange rate, were some of the best moments in university.

This study was partially funded by Scientific and Technological Research Council of Turkey (TUBİTAK) under grant numbers 118Z688 and 219Z213.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

ABBREVIATIONS

AKR1B10: Aldo-Keto Reductase B-10 Member

CCLE: Cancer Cell Line Encyclopedia

DFS: Disease-Free Survival

GSEA: Gene Set Enrichment Analysis

LASSO: Least Absolute Shrinkage and Selection Operator

LIHC: Liver Hepatocellular Carcinoma

OS: Overall Survival

PPP: Pentose Phosphate Pathway

RPPA: Reverse Phase Protein Assay

TCGA: The Cancer Genome Atlas

CHAPTER 1

INTRODUCTION

1.1 Tumor Statistics for 2020

There are many definitions for cancer, but the simplest is uncontrolled cell division resulting in tumors that can spread to other organs in the advanced stages (1). According to World Health Organization's Global Cancer Observatory (GLOBOCAN) statistics, there were 19.3 million new cancer cases and nearly 10 million cancer-caused deaths around the world for the year 2020, making it the leading cause of death globally, surpassing cardiovascular diseases and stroke for the first time (2). In addition, the incidence of breast cancer also is currently at highest with 11.7% of the total burden, followed by lung (11.4%) and prostate (7.3%). When it comes to total mortality, lung cancer still has the highest mortality rate (18%) followed by liver (8.3%) and stomach (7.7%) (Figure 1.1).

According to the same study, cancer incidence is expected to increase to 28.5 million cases globally by 2040. It is also estimated that about 1 in 3 people will develop cancer sometimes during their lifetime (2). Cancer is a highly complex disease with more than 200 varieties diagnosed to date, each having different characteristics and requiring different treatment approaches. However, all of them can be roughly classified into three main categories depending on their tissue of origin. Carcinomas, comprising over 90% of all cancers, are solid tumors originating from epithelial cells. Sarcomas are solid tumors originating from connective tissues of bones and muscle. Finally, leukemia and lymphomas arise from white blood cells. It is also common practice to name the tumors according to their organs of origins (1).

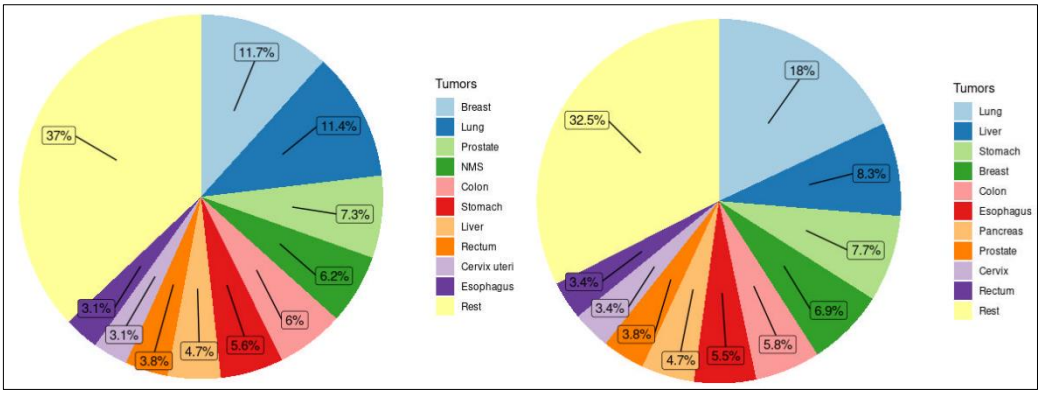

Figure 1.1. Global Incidence and Mortality for 10 Most Common Cancers in 2020. (Left panel) Percentage for the total burden and (right panel) percentage of mortality by cancer types. The data were retrieved from Sung *et al*. (2).

1.2 Basic Genetics of Cancer

According to their origin, cancers can be divided into two groups: sporadic and hereditary. Knudson first proposed the two-hit hypothesis according to which, for a sporadic cancer to develop and a normal cell to lose control of its normal division cycle, both alleles of a tumor suppressor gene need to be mutated. Given that such events are rare, sporadic tumors generally develop later in life. In the case of hereditary cancers, since at least one allele has already been inherited in the mutated form from one of the parents, receiving a second hit has a higher probability, so such patients develop cancer early in life (3).

Genetic changes can occur as a result of DNA replication errors which occur spontaneously in every cellular replication cycle since the DNA polymerase enzyme is not error-free. A second factor is mutations in DNA repair genes, which reduce the fidelity of the respective enzymes in fixing spontaneous DNA mutations that occur all the time. Another very important contributing factor is mutagens, which vary from solar UV radiation, to the food we consume, chemicals we are exposed to at work or the places we live in. Yet another important factor causing mutations and

even cellular transformation is viruses such as Human Papilloma or Sarcoma viruses, or hepatitis B and C viruses in the liver (1).

According to the current understanding, cancer is widely accepted as a genetic disease arising from mutational changes of the genome. High-throughput sequencing studies have shown cancers to have multiple genetic mutations such as single nucleotide changes leading to missense and nonsense mutations, multiple nucleotide deletions and rearrangements, large chromosomal translocations, gain or loss of whole chromosomal regions, as well as specific gene amplifications or deletions. In addition, cellular epigenetic profile such as methylation and acetylation are changed, as well as expression of long and short non-coding RNAs such as lncRNAs, miRNAs and snoRNAs involved in silencing of certain genes or overall chromosomal structure maintenance, among other functions (4).

Three types of genes are essential for controlling cell division; oncogenes, which are activated in cancer and drive cell division; tumor-suppressor genes, which serve as brakes on cell proliferation and have to be inactivated for a normal cell to be transformed into a cancerous one; and stability genes which are involved in maintenance of genomic integrity and also have to be inactivated for cell transformation to occur (1). Activation of oncogenes can occur either by gain-of-function mutations, epigenetic activation generally in the form of hypomethylation, or gene fusions resulting from chromosomal rearrangement, or insertion of viral expression elements near their promoters. Tumor suppressors and stability genes are inactivated from deleterious mutations or epigenetic silencing (5).

1.3 Hallmarks of Cancer

Hallmarks of cancer were first proposed by Hanahan & Weinberg (6) in an attempt to summarize the outcome of cancer research from different disciplines and identify common biological capabilities which sustain tumor growth, proliferation and

metastasis. These functional capabilities are proposed to be acquired by cancer cells via different mechanisms in various time points through the course of tumorigenesis allowing the cell to survive, proliferate and disseminate (7). As the understanding of the disease's pathology and molecular biology grew, they were expanded beyond the six original hallmarks (6), namely (a) sustaining proliferative signaling, (b) evading growth suppressors, (c) resisting cell death, (d) inducing angiogenesis, (e) enabling replicative mortality and (f) activating invasion and metastasis. Two enabling hallmarks which make possible the acquisition of the above six hallmarks are (g) genome instability and mutation, and (h) tumor-promoting inflammation. Finally, two emerging hallmarks of (i) deregulating cellular energetics and (j) avoiding immune destruction were also added to the list, making them a total of 10 hallmarks (7). Since the central topic of this study is tumor metabolism, a more detailed overview of this particular topic is provided in the section below.

1.4 Tumor Metabolic Rewiring

Reprogramming of energy metabolism with a shift from the more energy efficient oxidative phosphorylation (OxPhos) to fermentation in the presence of oxygen (also known as aerobic glycolysis or Warburg Effect) was first proposed and experimentally shown by Otto Warburg in 1926, unleashing a new era in the biochemical characterization of tumor cells with a focus on mitochondrial function (8,9). According to the original idea of Warburg, cancers arise from a gradual and irreversible respiratory impairment, and only tumor cells capable of increased aerobic glycolysis to compensate for OxPhos would be able to replicate and form cancers, while others would die as a result of insufficient energy (10).

The idea of impaired respiration in tumor cells was supported by four main quantitative variables: high glycolysis/respiration ratio, low absolute oxygen consumption, uncoupled or inefficient respiration, and low succinate oxidative response (11). With better understanding and characterization of glycolytic

processes and improvement in experimental procedures, aided also by the ability to obtain pure tumor cells and grow them in serum culture while carrying out the above-mentioned measurements, the distinction in tumor and normal cells became even sharper. In addition, isotope tracing experiments showed that normal cells produced at least twice as much labeled $^{14}CO_2$ as compared to tumor cells (11). However, the idea had its own opponents, the most outspoken of whom was Sidney Weinhouse who refuted it using a set of diverse and not well-controlled data to reject the hypothesis as too simplistic to explain a complex disease such as cancer (12).

Eventually Weinhouse won the argument partially aided by the death of Warburg in 1970 and because of his recognition of the importance of genetics (13). With improvements in genetic methods such as cloning and sequencing, focus shifted towards genetic mechanisms of cancer, leading to the proposal of "Somatic Mutation Theory of Carcinogenesis" according to which cancer is a genetic disease that advances with increased mutation burden (14). The discovery of oncogenes (15), proposal of the two-hit hypothesis by Knudson (3), the stepwise model of colorectal cancer development by Vogelstein (16) as well as major improvements in sequencing technologies after 1990s completely overshadowed the more cumbersome biochemical studies of tumorigenesis. These developments lead to the idea of finding a set of consistently recurring mutations in oncogenes and/or tumor suppressor genes and develop drugs to target them as a means of eradicating tumors just like the eradication of other diseases by means of antibiotics (17).

In order to achieve this aim, in 2005 a new idea was proposed to use new next-generation sequencing technologies (NGS) not only for whole tumor genome sequencing, but to also quantify their transcriptome, methylome and proteome as well, and for a fraction of patients their adjacent normal tissues were included to make it possible to compare deregulated signaling pathways and find druggable targets. This was called The Cancer Genome Atlas (TCGA) network and its first outcome was the study of human glioblastoma (18). In total, tumors from over

11,000 patients in addition to a fraction of normal adjacent tissues were sequenced and their gene mutations, genomic rearrangements, mRNA and miRNA expression, methylation patterns and a targeted number of important signaling proteins were characterized. The data were also made public in various format to be harnessed by researchers all over the world (https://www.cancer.gov/).

Despite the high hopes in this sheer amount of information made available by massive amounts of genomic, transcriptomic and methylation data, soon it was realized that cancers were highly heterogeneous; moreover, the same tumor in the same patient showed high degrees of heterogeneity, making it difficult to find a small set of targetable mutations or signaling pathways (19,20). One of the success stories of this approach was the development of Gleevec (Imatinib), which targets a very rare tumor known as chronic myelogenous leukemia (CML). However, Imatinib turned out to be an exception as other similar drugs targeting important signaling pathways and protein kinases turned out to be too toxic with no real long-term benefits for the patients (21). This is mostly because tumors, due to their heterogenic nature can quickly develop resistance against these drugs by undergoing clonal expansion, reducing the effectiveness of the drugs along with the return of a more aggressive form of the tumor that is frequently lethal for the patient.

Over 250 oncogenes and 700 tumor suppressor genes involved in almost every cellular function and signaling pathway have been catalogued and characterized to various degrees. A large proportion of these proteins play essential roles in controlling metabolic flux through key metabolic pathways that favor the growth and proliferation of tumors (22). In addition, anaerobic glycolysis originally proposed by Warburg is recognized as probably the only common process among all known tumors and has turned out to be essential for tumor detection by fluorodeoxyglucose positron emission tomography (FDG-PET) (23). Recognizing the importance of the Warburg effect, a new idea has started to become popular among researchers about

potential roles of oncogenes in controlling metabolic reconstitution necessary for tumor growth and proliferation (24,25).

Initially, large-scale studies harvesting microarray data supported this idea and a modified version of Warburg's original theory was proposed according to which metabolic rewiring of tumors under the control of oncogenes is an important part of tumor survival and proliferation (26). Other studies harvesting the information from TCGA datasets showed that metabolic genes (27) and transcription factors controlling their expression (28) undergo major changes in tumors to maximize their survival and growth capabilities, opening a new page in cancer research and expanding the original ideas of Warburg to include other important metabolic processes beyond glycolysis, such as amino acid, lipid and vitamin metabolism (29) in addition to the well-studied processes of nucleotide biosynthesis which have been used as drug targets for a long time.

Recent research has once again put mitochondria at the central stage of tumor metabolism, albeit not for bioenergetics but because of the central role of tricarboxylic acid cycle (TCA) and anaplerotic reactions. The metabolism of glutamate, which is the most abundant amino acid in circulation and can be used for energy production via TCA as well as biosynthesis of other macromolecules in the cell has taken the center stage (30,31). In addition, mitochondria have other well-known metabolic functions for cell growth and survival which had surprisingly been neglected until recently; it is the organelle from which acetyl-CoA, the precursor of fatty acid biosynthesis as well as histone acetylation is generated (32). Mitochondria provide cells with ample amounts of NADPH, which is essential for biosynthetic reactions and mitigation of oxidative stress via ROS (33), and it is the site of synthesis of important oncometabolites such as 2-hydroxyglutarate, succinate and fumarate as a result of mutations in certain TCA enzymes. These oncometabolites have important tumorigenic functions in hypoxia and epigenetic changes, but at the same time they show great potential for tumor screening and early detection (34,35).

1.5 Liver Cancer

1.5.1 Global Incidence, Mortality and Risk Factors

Liver cancer by incidence is the sixth most common tumor in the world with 906,000 new cases and second deadliest with over 830,000 (Figure 1.1). Moreover, this tumor is disproportionally more common in males compared to females with 2-3 times higher rates for both incidence and mortality. If we look at the data closely, liver cancer is the fifth most common tumor among males with 6.3% of total burden and tenth among females with only 3% in 2020. Despite the lower incidence, by mortality, it was the second deadliest in males with 10.5% and sixth in females with 5.7% of the total burden (2). Moreover, the disease shows heterogenic distribution and is more common in developing countries compared to developed ones. It is also the most common tumor among men in Eastern and Southeastern Asia as well as Northern and Western Africa, and its mortality rates are the highest in these four regions (2).

Liver cancer is not a single tumor type but a collection of subtypes, the most common being liver hepatocellular carcinoma (LIHC, also known as hepatocellular carcinoma, HCC), which accounts for nearly 80% of the total cases followed by intrahepatic cholangiocarcinoma with 15% and a number of other rare types accounting for less than 5% of the total burden (2,36). If the tumor is detected at an early stage it can easily be treated by surgical resection, however by the time it is diagnosed, less than 20% of the patients are eligible for this approach and as a result, the 5-year survival of liver cancer patients is less than 10%, while longer-term survival rate drops to 3%-5% (36).

There are a number of risk factors associated LIHC, most of them leading first to chronic infection and cirrhosis, and ultimately LIHC. Studies have shown that over 80% of LIHC cases have cirrhosis as a preexisting condition. Risk factors include

hepatitis B (HBV) and C (HCV) viruses, heavy alcohol consumption, food contaminated or enriched in aflatoxins, high body-mass index (BMI), type-2 diabetes, and smoking (37). As a result of successful vaccination campaigns against HBV and HCV, the number of hepatitis-related cases have dramatically dropped in many developing countries and now the increase in cases is mostly attributed to the non-viral causes mentioned above. One such important factor is nonalcoholic fatty liver disease (NAFLD), which is becoming prevalent especially in western countries due to increasing rates of obesity and associated metabolic syndrome. This has been correlated with the recent surge in the number of LIHC in developed countries, although their association is being disputed due to lack of supporting population-level studies (38).

1.5.2 Molecular Biology and Genetics of LIHC

Recent whole-genome sequencing studies have revealed a lot of important information about single nucleotide polymorphisms (SNPs) as well as major genomic rearrangements such as deletions and amplifications that take place during liver tumorigenesis (39). However, implementing these approaches to LIHC has not been as straightforward as in other tumors because of high stromal composition and normal cells included in the resected samples, thus confounding the results and making their replication difficult. In addition, as mentioned previously LIHC is generally the result of chronic conditions such as cirrhosis, hepatitis due to viral infection, metabolic syndrome and fibrosis, making the distinction between normal and tumor tissue challenging and blurring the stages from normal tissue to dysplastic lesions to full-blown LIHC (40).

In an attempt to understand genomic changes and correlate them with disease etiology to better classify and treat LIHC tumors, a large meta-analysis including 31 comparative genomic hybridization (CGH) studies for a total of 785 LIHC and 30 premalignant dysplastic nodules was carried out (41). The study showed

amplifications in 17q, 6p, 8q and 1q, varying from 22.2-57.1% of the cases. On the other hand, deletions in 13q, 17p, 4q, 16q and 8p were present among 26.2-38% of the patients analyzed. Some of these alterations correlated with the presence of HBV or differentiation grade of LIHC, while others were enriched in early stages of dysplasia. In addition, both amplifications and deletions were seen in the chromosomal arms on which well-established oncogenes and tumor suppressor genes such as *MYC* and *RB1* are located. Importantly, the study also showed that these regions also encode other suspected oncogenes that modulate the WNT pathway such as *AXIN2*, *WISP1*, and *FZD3* (41).

In similar studies based on integration of genomic data from various cancers, researchers have discovered 12 essential pathways that drive tumorigenesis by determining cell survival, cell fate and genomic maintenance (42). Based on these findings, LIHC is currently being considered as a multistep process that involves genetic and epigenetic changes, mainly affecting signaling pathways controlled by chromatin modifications and altered canonical pathways of β-catenin, TP53, Wnt, epidermal growth factor (EGFR) and Myc (42). These results were reconfirmed and expanded by the most comprehensive study on LIHC to date carried out by the TCGA network which among others showed that a number of genes essential for metabolic reprogramming such as Albumin (ALB), Apolipoprotein B (APOB), and Carbamoyl-Phosphate Synthase 1 (CPS1) which carries out the rate-limiting step of urea cycle were downregulated by either hypermethylation or mutation. In addition, the same study identified important therapeutic targets in the Wnt pathway such as MDM4, MET, IDH1 and TERT among others. Finally, the authors reported the importance of immune checkpoint proteins such as CTLA4, PD1, and PDL1 for targeted immunotherapeutic approaches in the future (39).

1.5.3 Diagnosis, Staging and Treatment of LIHC

Early detection of liver tumor is a major challenge as the disease goes through a long evolution from liver injury or chronic infection to cirrhosis and eventually LIHC. The tumor is generally detected at late stages and the survival of patients is very short. However, improvements in computer tomography (CT) scans and introduction of machine learning algorithms are making the process of tumor detection and image segmentation better (43). Recently, incorporation of the latest developments in deep neural networks with Gaussian mixture methods (GMM) trained on CT images increased tumor detection accuracy above 98% and they hold great promisses not only in aiding radiologists to detect LIHC cases as early as possible, but also to better understand the real growth and spread of the tumor in the organ (44).

Heterogeneity of liver tumors is a challenge for drug targeting as well. This is also compounded by the natural function of the liver as the center of metabolism where numerous chemicals are metabolized daily, including drugs (45). As a result, the best treatment options currently are resection by surgery and transplantation. For early stage tumors, radiotherapy and ablation have been shown to be very effective, however early detection in generally difficult in liver cancer (36). Adjuvant therapy options such as Sorefanib (46) and Brivanib (47) have not shown any significant difference in disease-free (DFS) or overall survival (OS) while acyclic retinoid decreased recurrence rate from 49% to 27% within a 3 years follow-up period (48). Many tyrosine kinase inhibitors that target vascularization have already undergone phase II and III clinical trial, however the extension of patient OS has been between 2-4 months while toxicity very high (49–53). Combination therapies have also been tried but to no avail as toxicity turned out to be a major preventative factor (54,55). After immunotherapy showed very promising results in other tumors such as leukemia, the feasibility of checkpoint blockage started to also be explored in LIHC. In a small pilot study of 20 patients with advanced stage of LIHC, anti-CTLA4 monoclonal antibody tremelimumab was used and the results were very encouraging

with patients having a median DFS of 6.5 months (56). Another similar monoclonal antibody which inhibits immune checkpoint signaling is the PD1 blocker is nivolumab. A phase II multi-center study showed nivolumab to be safe and the initial results of the ongoing trial are very promising (57).

The information provided in this section was not intended to be an exhaustive coverage of the available literature on all different therapeutic approaches in clinical use and/or under investigation against LIHC. However, they clearly show that most of the approaches do not provide any significant advantage to the patients, and even those few months of OS extension are generally accompanied by sometimes major side effects due to drug or treatment toxicity. As a result, there is an urgent need for better and more enduring approaches other than kinase inhibitors to improve both prognosis and treatment of LIHC.

1.6 AKR Genes and their Role in Tumorigenesis

1.6.1 General Characteristics of AKR Genes

AKR protein family has an ancient evolutionary history and all its members share a conserved $(\beta/\alpha)_8$ barrel plus a pyridine nucleotide binding pocket (58). There are currently over 150 genes belonging to this family distributed among three domains of life, and while some of them have been characterized biochemically, others are simply discovered by sequence analysis. Studies indicate it is highly likely that all these members originated by duplication events and were modified for different substrates through a long evolutionary history. Analysis of sequence homology have divided AKRs into 15 families with more than 40% sequence homology among each other, while between members of subfamilies the sequence homology is over 70% while structural homology is above 90% (59). Their nomenclature is very straightforward, starting with AKR followed by a number representing family (eg.

AKR1), then a letter representing subfamily (eg. AKR1B) and finally another number representing the member within the subfamily (eg. AKR1B10).

According to a report based on NCBI database search, a total of 58 actual and potential members of the AKR family were identified in the human genome (60). Among them, 15 members represented by three families have been characterized and classified as AKR1, AKR6 and AKR7. Furthermore, AKR1 is divided into 5 subfamilies, namely A (1 member: AKR1A1), B (3 members: AKR1B1, AKR1B10 and AKR1B15), C (4 members: AKR1C1, AKR1C2, AKR1C3 and AKR1C4), D (1 member: AKR1D1) and E (1 member: AKR1E2). AKR6 has 3 members (AKR6A3, AKR6A5, AKR6A9), and finally AKR7 has two members (AKR7A2 and AKR7A3). The remaining members were either pseudogenes, AKR genes predicted from sequence, or other proteins showing sequence and potentially functional similarity to AKRs (60). Our main focus in this study is AKR1B10 and some members of AKR1C family.

Aldo-Keto-Reductase family 1, member 10 (AKR1B10) enzyme was originally named as aldose reductase-like-1 (ARL-1) (61,62) due to its sequence homology with AKR1B1 member of the same subfamily, the first to be characterized and implicated in diabetic pathophysiology (58). AKR1B10 is a cytosolic enzyme showing 71% sequence identity to AKR1B1 and utilizes NADPH as a cofactor to reduce a variety of substrates such as retinaldehydes, lipid peroxidation byproducts, as well as xenobiotics. AKR1B1 on the other hand shows higher preference for glucose, which it reduces to sorbitol, a pathway that may have very important implications in cancer metabolism (63). Studies have shown that during hyperglycemia, AKR1B1 can convert up to 30% of total glucose into sorbitol, while in normal condition it reduces only about 3% (64). Additionally, AKR1B1 is ubiquitously expressed among multiple tissues while AKR1B10 is highly expressed in gastrointestinal tract in addition to thymus, liver and kidney, albeit at lower levels in the last three organs. Both genes are located next to each other on chromosome

7q32-33 bands and each contains 10 exons. The protein products of both genes consist of 316 amino acids and there are no reports of alternatively-spliced variants to date (61,62). AKR1B15 on the other hand was characterized very recently in a study which showed that the gene is located on the same chromosome as the other two AKR1B members, it has over 90% sequence homology with AKR1B10, and the enzyme uses NADPH as a cofactor. However, based on limited enzymatic studies, it was shown to act as a steroid reductase, to be predominantly localized in mitochondria and highly expressed in adipose tissue, testis and placenta (65).

Similar to AKR1Bs, AKR1Cs also utilize NADPH as a cofactor for their enzymatic reactions but they are clustered together on chromosome 10, p15-p14 bands (66). They share much higher sequence homology, ranging between 84%-98% (67), and different from AKR1B subfamily, their main function is steroid metabolism, especially phase I biosynthetic reactions (68). AKR1C1 and AKR1C2 are expressed in many tissues, while AKR1C3 and AKR1C4 are expressed almost exclusively in prostate and liver, respectively. More specifically, AKR1C enzymes are involved in androgen biosynthesis, particularly 5a-dihydrotestosterone (5a-DHT) and its inactivation, metabolism of estrogen and progesterone, neurosteroids, bile acid synthesis (particularly AKR1C4 in liver), and side chain of steroids to make them prone to conjugation by phase II enzymes (69).

1.6.2 Enzymatic Actions and Tumorigenic Effects of AKR1 Family

AKR proteins share a common $(\beta/\alpha)_8$ barrel and a nucleotide- binding pocket which is occupied preferentially by NADPH. Members of the AKR1 family are monomeric, while members of other families form higher quaternary structures, more particularly dimeric (AKR7) and tetrameric (AKR6) through certain residue interactions found in specific loops (59). The 3-dimensional structure of AKR1B10 bound to $NADP^+$ is shown in Figure 2.2.

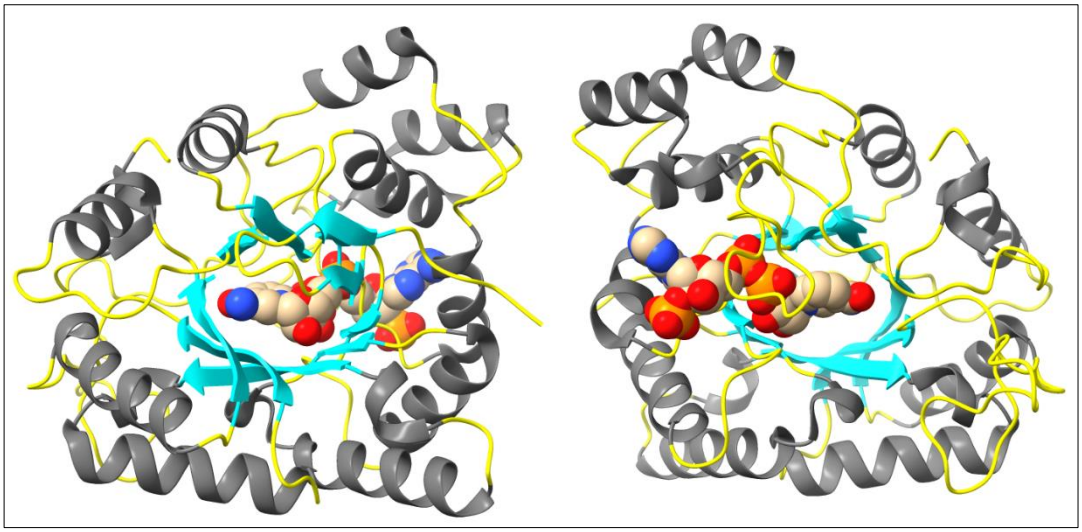

Figure 1.2 Three-Dimensional Structure of AKR1B10

The two images represent the front (left panel) and rear (right panel) view of AKR1B10 protein and they are colored according to secondary structures. Dark grey shows 8 alpha-helixes, cyan the 8 betta-strands and yellow the loops connecting them. $NADP^+$ is shown in ball-and-stick mode. The structure with PDB ID 4GQG was downloaded from protein databank (58) and images were generated with ChimeraX (70).

Although there are minor differences in reaction mechanisms of each AKR1 member, overall they consist of two steps: first, a hydride ion is transferred from NADPH to the substrate, forming a ternary complex of enzyme-substrate-cofactor. Next, a hydrogen proton from the solvent attacks and releases the aldol form of the substrate and the oxidized form of $NADP^+$ (59). The substrates of AKRs include a wide range of chemicals that contain aldehyde and ketone moieties that are reduced to alcohols, making them indispensable for detoxification of many carbonyl compounds (AKR1B10) that are either produced as byproducts of normal cellular processes or cellular stress, chemical transformations which could be activation or deactivation of steroid hormones (AKR1C), or detoxification of xenobiotic compounds, many of them used in cancer chemotherapy, making them available to phase II conjugation processes for clearance from the body (AKR1B and AKR1C)

(71). Many *in vitro* enzymatic experiments with different substrates are available in some excellent studies (61,72–75).

Since AKR enzymes in general and AKR1 family in particular require NADPH for their enzymatic actions, they may have broad but yet unexplored implications for cellular physiology. NADPH is synthesized through many metabolic pathways in the cell, first and foremost by pentose phosphate pathway (PPP) which branches from glycolysis, but depending on cell type and context, other pathways such as one-carbon cycle and anaplerotic reactions can also generate ample amounts of the cofactor (33). Cells maintain a high NADPH/NADP ratio even during glucose shortage partially to cope with oxidative stress among many other essential cellular processes (76). Thus, overexpression of AKR proteins, in addition to their detoxification functions, may also tip the balance of NADPH/NADP which may potentially affect certain cellular processes.

AKR1s, in addition to general detoxification reactions have also been found to be important in retinoid signaling which is essential for cellular differentiation. The enzymes play an important role at the junction where retinal is either oxidized by aldehyde dehydrogenases (ALDHs) to retinoic acid and lead to cellular differentiation, or reduced by AKR1B10/C1/C3 to retinol, in this way depriving the cell of retinoic acid. This balance is thought to be determined by the availability of the aforementioned enzymes (77). Another important process where AKR1B10 is involved is in protein prenylation through the mevalonate pathway. Enzyme kinetic studies showed low K_m and high catalytic efficiency of AKR1B10 compared to AKR1B1 towards Farnesal (FAL) and Geranylgeranial (GGAL), converting them into their alcohol forms (78). Similar to retinoids, aldehyde FAL and GGAL are intermediates that can be either oxidized by alcohol dehydrogenases (ADH) to their carboxylic forms which among others induce apoptosis, or reduced by AKR1B10 into alcoholic products FOH and GGOH, which are then phosphorylated to pyrophosphate forms and used to prenylate cellular proteins mostly involved in

signal transduction such as G-proteins regulating MAP kinase pathways and cellular proliferation (79).

An additional role of AKR1B10 was found when Ma et al. (2008) reported that AKR1B10 blocked degradation of acetyl-CoA carboxylase α (ACACA), the rate limiting enzyme for de novo fatty acid synthesis, thus enhancing tumorogenicity of breast cancer cells (80). These authors showed that AKR1B10 could physically interact with ACACA and block ubiquitin-dependent degradation. This could be a novel role of AKR1B10 in addition to its conventional enzymatic function. Last but not least, AKR1B10 is involved in cancer chemoresistance. Studies have shown that AKR1B10 reduces the antiemetic 5-HT3 receptor antagonist dolasetron, anticancer drugs oracin and daunorubicin, as well as tobacco carcinogenic compounds such as NNK. These actions happen in cancers with high expression of AKR1B10 such as liver and lung where AKR1B10 inactivates the drugs by either reducing them into less harmful forms, or by reducing and detoxifying the highly reactive carbonyl products generated by ROS after chemotherapy, which are normally expected to damage tumor cells and force them to undergo apoptosis (79). In both ways, AKR1B10 and other members of AKR1 family have been shown to play crucial role in cancer resistance to chemotherapy; therefore, agents that block them during such treatments are still under investigation (81).

The expression of AKR1B10 shows a mixed pattern with up- or downregulation in various tumors. In some tumors expression of AKR1B10 is associated with poor prognosis, while in others with better prognosis for tumor patients. Various studies have shown that expression of AKR1B10 is downregulated while that of AKR1B1 is upregulated in colorectal cancer. Their expression was combined into a gene signature that could stratify patients into low and high-risk groups (82,83). Another study showed that AKR1B10 expression is regulated by p53 in colon and its suppression inhibits apoptosis, while its overexpression inhibited cellular proliferation by a yet unknown mechanism (84). However, Yan et al. (2007) found

that silencing AKR1B10 by siRNA in HCT-8 colon cancer cell line reduced proliferation of cells treated with acrolein and crotonaldehyde, two toxic ketogenic compounds (85). These seemingly contradictory results can be explained by the enzymatic function of AKR1B10 which is reduction of toxic aldose compounds into less harmful alcoholic forms. Gastrointestinal tract in general, and colon in particular are under constant pressure of organic toxic compounds coming either from the food or byproducts of gut microbiome, so downregulation of AKR1s would expose the cells to attacks, causing mutation and cancer. On the other hand, during chemotherapy AKR1s acts again in the same way, reducing the drugs into harmless compounds, making the cells resistant (86).

The expression of AKR1B10 has also been reported to be upregulated in many tumor types, first and foremost in LIHC from which the gene was first isolated and characterized (61). Thus, in an early study using tissue microarray with immunohistochemistry staining, the elevated status of AKR1B10 in most of the tumors as compared to normal adjacent tissues was confirmed, but higher AKR1B10 expression was associated with better DFS and OS (87). In contrast, according to another report using the same approach, high AKR1B10 levels were associated with poor OS and DFS and this was linked to AKR1B10 metabolizing sphingosine 1-phosphate (S1P) which in turn induced cellular proliferation (88). In agreement with the last study, a large multi-center study found elevated levels of AKR1B10 protein in serum of early and late stage LIHC patients. Additionally, the same report showed the potential of AKR1B10 as a biomarker for early and late-stage tumor detection in AFP-positive and negative LIHC patients. However, when both proteins were combined into a single risk signature, sensitivity and specificity were better than each of them individually (89).

Another recent study put the expression AKR1B10 under the control of IRAK1, a member of NF-κB pathway and found AP1 binding site on its promoter. Higher expression of IRAK1 was associated with poor OS, and this observation was linked

to increased drug resistance and cellular stemness in LIHC cell lines (90). These results show that there is a contradiction among different clinical studies regarding the status of AKR1B10 and its potential as a biomarker for LIHC. These contradictions were addressed in a review analyzing 9 clinical studies regarding the association between AKR1B10 expression and LIHC patients' clinical outcome, and according to the authors, the differences can be explained by early versus advanced LIHC stages as well as the presence of HBV, HCV, Cirrhosis and other yet unknown factors involved in initiation and advancement of LIHC (91).

Enhanced expression and specific activity of AKR1B10 and AKR1B1 has been reported in all stages of breast cancer as compared to normal adjacent tissue, both before and after surgery and this was associated with drug resistance. To prove that claim, the same study showed that expression of these two enzymes was induced by proteasome inhibitor bortezomib in HT-29 breast cancer cell line in vitro (92). Another study also showed that expression of AKR1B10 was elevated in HER2+ and ER- breast cancers, however its overexpression does not affect tumor behavior in primary sites, but it is associated with increased metastatic relapse in secondary sites. In addition, in vivo studies showed that tumor cells with high AKR1B10 expression shifted their dependency from glycolysis to fatty acid oxidation, a phenomenon that can explain the association of AKR1B10 expression and metastatic relapse. Finally, the same study showed that AKR1B10 protected cells from oxidative stress generated by fatty acid oxidation (93). High expression of AKR1B10 was also associated with low DFS while showing no association with OS in lung adenocarcinoma (94).

1.7 NADPH Biosynthetic Pathways in Eukaryotic Cell

There are three main sources of NADPH in the eukaryotic cell: PPP, one-carbon cycle (OCC) also known as the folic acid cycle, and anaplerotic reactions, in particular Malate Dehydrogenase or Malic Enzyme 1 (ME1) but also to a certain

extents Isocitrate Dehydrogenase 1 (IDH1) in the cytosol and IDH2 in the mitochondria (Figure 1.3).

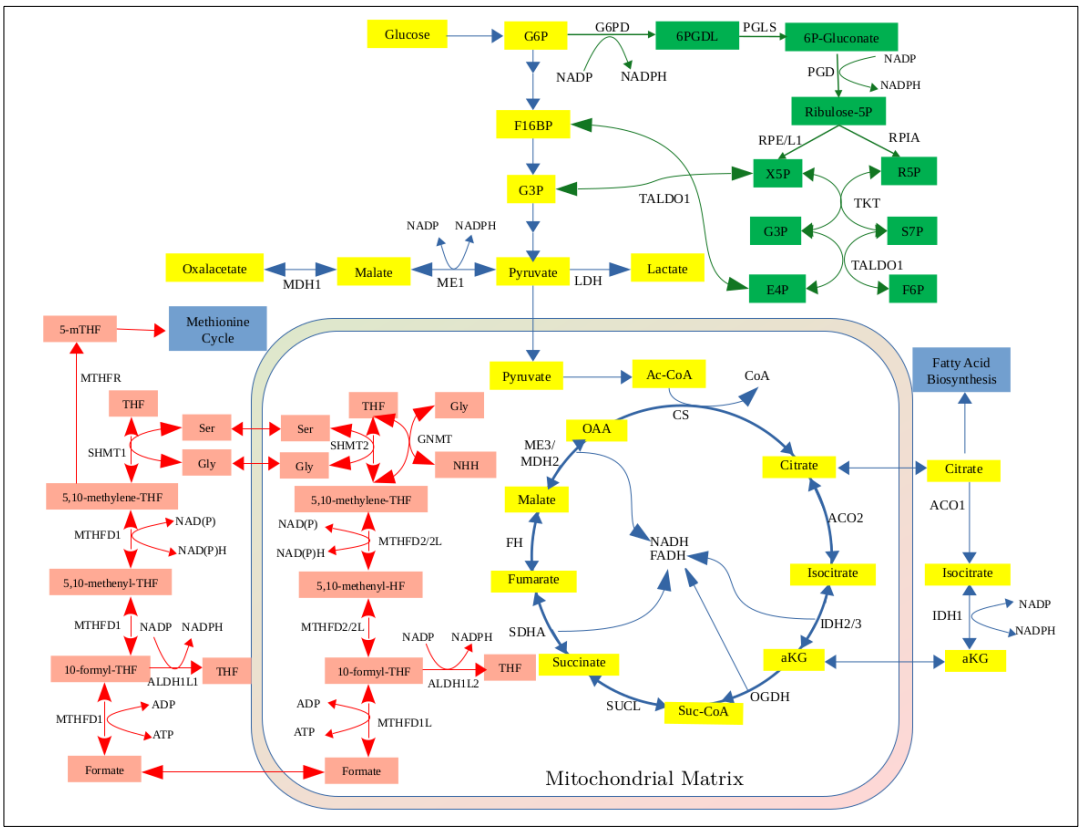

Figure 1.3 NADPH Synthetic Pathways.

Yellow color shows glycolysis, TCA and anaplerotic reactions, green color shows PPP, red shows OCC, and blue shows major pathways the reactions lead to.

Metabolic flux studies have shown that PPP is the main contributor, but depending on cell type and context, OOC can sometimes take over (33). PPP consists of two branches, the oxidative branch which oxidizes glucose 6-phosphate to ribulose 6-phosphate and uses those electrons to reduce $NADP^+$ into NADPH in three enzymatic steps. Two of these steps include the enzymes Glucose 6-Phosphate Dehydrogenase (G6PD) and Phosphogluconate Dehydrogenase (PGD) that carry out the actual oxidation reactions and produce 2 molecules of NADPH. The other branch is called the non-oxidative branch which is carried out by TKT and TALDO1

enzymes; it uses glycolytic intermediates fructose 1,6-bisphosphate (F16BP) and glyceraldehyde 3-phosphate (G3P) to synthesize ribulose 6-phosphate leading to nucleotide biosynthesis, or vice versa. The activity of each branch is determined by cellular needs. Thus, during anabolism cells need to synthesize new lipids and nucleotides as well as to keep oxidative stress under check, and in this setting both the reducing and non-reducing arms of the pathway direct the metabolite flow towards the production of more NADPH and nucleotides. However, during the catabolic mode, the non-reducing arm works in reverse and funnels the metabolites into glycolysis for producing ATP. In this way the cell generates the necessary reducing power to keep oxidative stress under check and satisfies its energy needs (95).

Folate is an important element of one-carbon metabolism and its real biological significance has not been understood yet. Some of the cellular processes it is involved include nucleotide synthesis, gene expression regulation, cell division, and metabolism of amino acids, especially Serine and Glycine. Folate was recently shown to be a major contributor of the cellular NADPH pool, almost equal to PPP, depending on the physiological conditions of the cell (33). Briefly, the pathway takes place in the cytosol and mitochondria, with some enzymes having similar functions but located in different cellular compartments (Figure 1.3). By convention, cytosolic enzymes are labeled as 1 and mitochondrial enzymes as 2. The enzymes involved in reducing NADP to NADPH are methylenetetrahydrofolate dehydrogenase 1 (MTHFD1) and Aldehyde Dehydrogenase 1 Family Member L1 (ALDH1L1) in cytosol, and MTHFD2 and ALDH1L2 in the mitochondria.

The third main source of NADPH in the cell comes from the anaplerotic reactions of the TCA cycle, particularly from three reactions, one in the mitochondria and two in the cytosol (Figure 1.3). The first and most important one is cytosolic ME1 which oxidizes malate transferred from the mitochondria via the malate-aspartate shuttle into pyruvate by reducing one $NADP^+$ to produce NADPH. The second enzyme is

IDH1 which oxidizes Isocitrate into α-Ketoglutarate (αKG) in the cytosol by transferring the electrons to $NADP^+$ to produce NADPH, or vice versa. IDH1 enzyme is unique because when mutated, instead of converting Isocitrate into αKG, it reduces the latter into 2-Hydroxyglutarate (D-2HG), an oncometabolite that is found commonly in gliomas and myeloid leukemia. D-2HG is known to inhibit the enzymatic actions of many αKG-dependent dioxygenases, such as histone demethylases, thus causing major epigenetic changes in cells harboring such mutations (96). The final NADP-reducing enzyme is IDH2, the counterpart of IDH1 functioning in mitochondria which oxidizes Isocitrate into αKG during TCA and transfers the electrons to $NADP^+$ instead of NAD^+, which is the preferred cofactor of IDH3 and it has important roles in mitigating mitochondrial ROS (97).

1.8 Transcriptomics as Surrogate for Determining Cellular Metabolic State

Given the complexity of regulatory mechanisms used by cells to regulate enzyme expression and stability, as well as the flux through metabolic pathways, it may seem unfeasible to try to understand metabolism by means of transcriptomics. However, recent studies on proteomics have established a strong correlation between gene transcription and protein abundance in many cancers, especially the liver (98). It is a well-known fact that metabolic flux depends on multiple factors, such as the abundance of the metabolite itself, the availability and kinetics of one and/or multiple enzymes, competition among many enzymes for the same substrate where there is pathway branching as well as enzyme inhibition by various negative feedback mechanisms. This is further compounded by post-translational modifications of the enzymes and SNPs, which can drastically change their kinetics (99). However, recent integrated studies on transcriptomics and metabolomics for breast cancer have shown strong evidence that metabolite abundance and therefore metabolic flux is strongly correlated with transcriptional abundance in tumors (100,101). Additionally, using transcriptomics to understand metabolomic profile of cancers is

an easy and effective way when compared to the more tedious, complicated, expensive, and still not well-established methods of capturing, defining and quantifying all the metabolites in a cell or tissue (102,103).

Several studies based on transcriptomics either from microarray or RNA-sequencing have already used transcriptomics to understand the metabolic state of tumors. In one of the earliest studies of this nature, metabolic gene expression differences between 22 different tumors and their matched normal tissues were compared and it was found that different tumors overexpress different isoforms of metabolic enzymes, especially glycolytic ones. Moreover, despite changes in metabolic gene expression, overall metabolic profiles of tumors were more similar to their tissue of origin than to other tumors (104). In another more recent study, TCGA data were utilized to compare metabolic gene changes between different tumors, and major dysregulation in mitochondrial genes involved in oxidative phosphorylation were detected. This state was correlated with poor prognosis and upregulation of EMT genes (27), a case that lends support to the idea of respiration injury in tumors.

Another similar study also used TCGA data to compare the expression of genes comprising 7 major metabolic pathways between 33 tumor types and classify them accordingly. It was found that increased beta-oxidation was associated with better patient prognosis, whereas elevated expression of genes involved in vitamin, carbohydrate and nucleotide metabolism was associated with poor prognosis. The remaining two pathways of energy integration and amino acid metabolism had mixed results and were dependent on tumor type (101). In one of the latest such studies, the expression of genes for 114 metabolic pathways from KEGG database were evaluated in 26 tumors and their adjacent normal samples. In addition to reporting pathway dysregulations that were either tumor-specific or common across many tumor types, changes in the expression of a number of important transcription factors called master metabolic transcription regulators (MMTRs) were also detected (102).

All these studies have established the accuracy and feasibility of transcriptomics as an accurate representation of the metabolic state of the cells. Based on these premises, we also used transcriptomics to study and interpret the changes in metabolic profile of carbohydrate, lipid and amino acid across multiple cancers and their adjacent normal tissues, as well as all metabolic the pathways in LIHC.

1.9 Aims, Hypothesis and Novelty

Cancer metabolism has gathered huge momentum in recent years because of its great potential in cancer management and therapy, and also because of the failure of conventional therapeutic approaches to deliver according to expectations. We started this study by manually curating an extensive set of enzymes as well as carbohydrate and amino acid transporters and compared the expression of these genes between cancer and matched normal tissues in 20 different tumor types We selected all enzymes that take part in carbohydrate, lipid, and amino acid metabolism, both catabolic and biosynthetic.

Additionally, all transporters for carbohydrate and amino acids in the cell, as well as some important mitochondrial channels were included. Finally, we also evaluated the expression of genes that are involved in anaplerotic reactions that are important for integrating these pathways together, and electron transport chain (ETC) in an attempt to understand whether mitochondrial ETC was deregulated in tumors. To our knowledge, this is the first study to specifically evaluate the expression of this specific set of genes and find important commonalities and differences between tumors. We also evaluated the prognostic value of some of these enzymes and transporters across different tumors.

The pan cancer transcriptome data indicated major deregulation of NADP/NADPH utilizing enzymes across many tumor types. We specifically selected LIHC as liver is the hub of metabolism and showed major deregulation in NADPH pathways. We

observed a strong and significant upregulation in AKR1B10 at very early stages of LIHC; this enzyme utilizes NADPH and is considered as a potential biomarker for poor prognosis. Using different *in silico* approaches, we evaluated the hypothesis that there exists an interaction between AKR1B10 and PPP, a major branch of central carbohydrate metabolism that reduces NADP to NADPH. Our data suggest that LIHC tumors with high AKR1B10 expression showed very strong association with PPP and other NADPH related pathways, along with enzymes of Phase II conjugation reactions. Thus, it can be envisaged that AKR1B10 high tumors can detoxify chemotherapeutic drugs and mitigate oxidative stress, which would otherwise slow their progress.

Finally, genes that were a part of a network with AKR1B10 as the hub was used to train a multi-gene prognostic risk signature and test it on external datasets using two different machine learning approaches. This was done in an attempt to better stratify LIHC patients across different datasets and we showed once again the importance of PPP as possible target for better patient stratification and tumor therapeutics.

To our knowledge, this is the first detailed study of the possible molecular mechanisms that increase the aggressiveness of $AKR1B10^{High}$ tumors. It is also the first to hypothesize and show a direct link between AKR1B10 and PPP using *in silico* tools. Finally, we also built a prognostic score that can be generalized across datasets, and although its accuracy is not as high as some already published in literature, its main advantage is its specificity as it is based on a single molecular pathway whose expression can be manipulated or targeted in the future.

CHAPTER 2

MATERIALS AND METHODS

This study was carried out on publicly available data from patients and cell lines. Also, all the software and functions to carry out the analysis are open source. The sections below will describe briefly the methods and parameters used in data mining, processing and statistical analysis to obtain the results presented in the next chapter.

2.1 Datasets

We used 24 publicly available datasets in this study, 20 TCGA tumor cohorts, 3 tumor microarrays and one CCLE. The TCGA cohorts and the number of samples for each of them are shown in Table 2.1. It should be noted that half of the samples represent tumors and the other half normal adjacent tissues.

GSE6764 contains 75 liver samples taken from 48 patients of various stages, starting from healthy liver tissue to advanced stages of liver tumor. Samples were collected in three different medical centers in New York (USA), Barcelona (Spain) and Milan (Italy). The transcriptome of these samples was quantified using Affymetrix HGU-133 Plus2 array platform (105). GSE76427 dataset contains 167 samples, 115 primary liver tumors and 52 adjacent healthy tissue, all collected in Singapore. The transcriptome of these samples was quantified using Illumina HumanHT-12 V4.0 expression bead chip (106). GSE14520 dataset contains 225 primary liver tumor samples obtained from Chinese and US citizens and the transcriptome was quantified by Affymterix HGU-133A array platform (107). In this study, only the normal samples from GSE6764 were used.

Table 2.1 Information Regarding TCGA Cohorts

Cohort ID	*Cohort Name*	*Sample Number*
BLCA	Bladder Urothelial Carcinoma	38
BRCA	Breast Invasive Carcinoma	224
CESC	Cervical and Endocervical Cancers	6
CHOL	Cholangiocarcinoma	18
COAD	Colon Adenocarcinoma	82
ESCA	Esophageal Carcinoma	16
HNSC	Head and Neck Squamous Cell Carcinoma	86
KICH	Kidney Chromophobe	46
KIRC	Kidney Renal Clear Cell Carcinoma	144
KIRP	Kidney Renal Papillary Cell Carcinoma	62
LIHC	Liver Hepatocellular Carcinoma	100
LUAD	Lung Adenocarcinoma	118
LUSC	Lung Squamous Cell Carcinoma	102
PAAD	Pancreatic Adenocarcinoma	8
PCPG	Pheochromocytoma and Paraganglioma	6
PRAD	Prostate Adenocarcinoma	104
READ	Rectum Adenocarcinoma	20
STAD	Stomach Adenocarcinoma	27
THCA	Thyroid Carcinoma	118
UCEC	Uterine Corpus Endometrial Carcinoma	70

2.2 RNA Sequencing Data Acquisition and Pre-processing

All TCGA RNA sequencing data were downloaded from TCGA repository by *TCGABiolink* package (108,109). In order to obtain better quality data, instead of using the older RSEM version of read counts which are available on firehose database (http://gdac.broadinstitute.org/), we downloaded the more recent and more accurate HTSeq read counts for each cohort (110). The code below shows the function for downloading HTSeq read counts for LIHC cohort.

```
CancerProject = "TCGA-LIHC"
queryDown = GDCquery(project = CancerProject,
                     data.category = "Transcriptome Profiling",
                     data.type = "Gene Expression Quantification",
                     workflow.type = "HTSeq – Counts")
GDCdownload(query = queryDown)
LIHCRnaseqSE = GDCprepare(query, save = TRUE,
                          save.filename = "TCGA_LIHC_gene_exp.rda",
                          summarizedExperiment = FALSE)
```

First, the cohort name is defined, which is first located on the cloud by *GDCquery()* function and downloaded on the machine by *GDCdownload()*. Then, the data is converted into table format by *GDCprepare()* function and saved into RData format (.rda) which is easy to upload into R software for downstream analysis. By changing the name of "CancerProject" according to the cohort IDs shown in table 3.1, the data for all cohorts can be downloaded for further processing. Next, clinical data for all samples was extracted and tumor to normal matching was done by patient id. In addition, all non-aligned read counts were removed from the expression matrix by detection and removal of non-ensembl IDs. At the end of this process, the results shown in table 3.1 were obtained. For LIHC on the other hand, we used all the tumor samples, not only those with normal matches. This was based on clinical data to remove all patients without survival follow-up, or more than one sequenced sample per patient, and we were eventually left with 364 primary LIHC tumors and 50 normal samples.

2.3 RNA-Sequencing Differential Gene Expression

For differential gene expression (DGE), *DESeq2* package was used according to instructions (111). Briefly, patients were divided into two groups, normal and tumor or $AKR1B10^{High}$ and $AKR1B10^{Low}$, or any other combination necessary, and a general linear model (GLM) was constructed with a single variable. Afterwards, all

probes with zero read counts were filtered out of the expression matrix to make computation easy and DGE was carried out. Statistical significance of the genes was calculated by Wald's test corrected for multiple hypothesis by Benjamini-Hochberg (BH) procedure referred to as FDR in the text (FDR < 0.05). However, the log2 fold change (LFC) varies according to our purpose and it is specified on every occasion. In addition, we also obtained the variance-stabilizing transformations (VST) values of each gene which produces data in the logarithmic scale of base 2 (log2) by removing the dependence of the variance on the mean. This is achieved normalizing the data with respect to the dataset size and normalization factors introduced in the GLM (112). In this way, VST transformation removes the bias introduced by adding a constant to deal with zero values which in contrast to microarray are very common in RNA sequencing. The VST values are used for all downstream analysis, such as boxplots showing gene expression between various conditions, clustering and heatmaps.

2.4 Methylation Analysis

Raw methylation signal files (.idat) were downloaded from TCGA database using *TCGABiolink* package. Afterwards, they were separated according to conditions explained in results chapter based on matches of patient ids between RNA sequencing and methylation. For sample pre-processing we used *minifi* package (113) and quantile normalization (114). All samples passed the quality criteria (mean-detection p-values < 0.05) and all the probes with no signal were removed. Differential methylation analysis (DMA) was carried out using *limma* package with M-values while beta-values are used later for visualization purposes. The way beta and M-values are calculated differs, and the common practice is to use M-values for DMA and beta values for visualization purposes (115).

2.5 Microarray Data Acquisition and DGE

There were no raw transcription quantification values for GSE76427 dataset, therefore the processed and normalized values were downloaded from gene expression omnibus (GEO) database using the *GEOquery* package from Bioconductor and log2-transformed. The same approach was used to download clinical data as well. The normalization quality of expression data was checked before they were used for further analysis. Some genes such as AKR1B10 and AKR1B15 which are important for this study had negative expression values as a result of normalization, therefore all the samples having negative values for these genes were removed when building the cox regression models. For GSE14520 and GSE6764 datasets, raw expression data were available on GEO, therefore they were downloaded with *GEOquery* package and normalized using robust multiarray averaging (RMA) method (116,117). Probe annotation for GSE14520 and GSE6764 were done using the *hgu133a.db* and *hgu133plus2.db* packages from Bioconductor, respectively.

GSE6764 was used for DGE analysis which was carried out using *limma* package (118). As will be shown in section 3.2, DGE analysis were carried out multiple times for this particular dataset, so each time the patients were divided into different groups based on the clinical data and differential expression matrices were designed accordingly. Afterwards, DGE was carried out by fitting a GLM to the data and the statistical significance for each probe was determined by t-statistic with BH correction (FDR < 0.05). The LFC cut-off for this dataset was 0.5 LFC between the conditions of interest.

2.6 CCLE RNA Sequencing and Drug Sensitivity Data

The pre-aligned, raw read counts in HTSeq format for more than a thousand cell lines were downloaded from the cancer cell lines encyclopedia (CCLE) database of

Broad Institute (https://portals.broadinstitute.org/ccle/data). Then, expression data for 25 liver tumor cell lines were selected and normalized with DESeq2 package using a GLM with a constant as variable since we were not interested in differential expression but only data normalization. The normalized and VST-transformed read counts were then used for further analysis. Metabolome data for CCLE were also downloaded from the same database and no processing was necessary as the data were already in normalized and clean format.

We also utilized the drug sensitivity analysis on liver tumor cell lines from Cancer Therapeutic Response Portal (CTRP) (http://portals.broadinstitute.org/ctrp/). Briefly, AKR1B10 gene and liver cell lines were selected and queried for correlation of drugs' area under the curve (AUC) values and AKR1B10 expression in the portal. Candidates were selected and their correlations re-confirmed by using CCLE RNA sequencing expression data and AUC values obtained from National Cancer Institute database (https://ocg.cancer.gov/ctd2-data-project/broad-institute-screening-dependencies-cancer-cell-lines-using-small-molecules-0). All correlation coefficients and their significance were calculated by Pearson correlation.

2.7 Reverse Phase Protein Array

Reverse Phase Protein Array (RPPA) is a recent technology developed to quantify a selected set of proteins across multiple samples and conditions simultaneously. It cannot quantify all peptides in the cell like mass spectrometric approaches, but depending on the purpose, the necessary antibodies are used to design chips and quantify proteins accordingly. It is especially useful for studying proteins involved in signal transduction and/or a selected set of biological pathways because it can detect both the protein and its modified version by post-translation modifications such as phosphorylation (119). We downloaded the pre-processed and normalized RPPA data for 184 LIHC samples from Firebrowser database (http://gdac.broadinstitute.org/) and matched them with RNA sequencing samples

according to unique patient ids and carried out differential expression analysis using *limma* package.

2.8 Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) is a sensitive method for detecting whole pathways enriched in one condition as compared to another. For this study, GSEA desktop application was downloaded from Broad Institute (http://www.gsea-msigdb.org/gsea/downloads.jsp) and it was run in Windows operating system. Expression and phenotype files formats and their contents were prepared in accordance with the software's user guide (https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideFrame.html). Expression file was composed of VST-transformed read counts and ensemble ids were used for gene annotation. All enrichments were run in 1000 permutations and different molecular signature databases (MsigDBs) were used, all of them explained in more detail in the relevant sections in Chapter 3.

2.9 Gene Ontology Terms Enrichment

For Gene Ontology (GO) term enrichment, various online servers such as David (https://david.ncifcrf.gov/tools.jsp), and Gprofiler (http://biit.cs.ut.ee/gprofiler/gost) were used, however their output format is difficult to work with. Instead, all the results shown in this study were the output of *clusterProfiler* package run on R (120). Overall, there was good agreement between the three approaches, but in addition to good graphical representation, *clusterProfiler* has the advantage of further trimming the number of GO output by removing GO terms with common genes. One special requirement of *clusterProfiler* is that it works only with ensemble gene ids to carry out enrichment analysis, and for RNA sequencing this was easy as all the genes were annotated with ensemble numbers. However, microarray annotation is based on probes which were converted into ensemble IDs by using genome-wide annotation

package *org.Hs.eg.db* from Bioconductor. All enrichments were carried out on genes passing the significance threshold (FDR < 0.05) and LFC > 1 for RNA-sequencing and LFC > 0.5 for microarray. We also checked for 3 ontology terms, biological process (BP) which enriches genes according to their overall biological function in the cell such as certain biological pathways, cellular component (CC) which enriches genes according to their locations in various cellular components such as organelles or plasma membrane, and molecular function (MF) which enriches genes by their activities such as certain biochemical reactions or cellular processes such as transport (http://geneontology.org/docs/ontology-documentation/). All the GO term enrichments in this study were done using the same stringent statistical criteria of p < 0.01 and its BH-correction (FDR) < 0.05. After trimming the GO output, a number of GO terms were selected and represented graphically using the *dropGO()* and *barplot()* functions from the same package, respectively.

2.10 Metabolic Enzymes Database

To date, there does not exist a comprehensive and consistent database representing all metabolic enzymes of the eukaryotic cell, but the numbers are changing according to enzymes' cellular function (121). In order to collect as many enzymes as we could for this study, an indirect approach was taken. First, using the *org.Hs.eg.db* package, all the ensemble gene IDs were converted into Enzyme Commission (EC) number, a unique ID for each enzyme annotating its main properties which are assigned according to the specifications of the International Union of Biochemistry and Molecular Biology (IUBMB). Next, all metabolic genes available in the KEGG database were downloaded and their entrez IDs were also converted into ensemble IDs, a unique gene annotation number that solves the confusion of multiple symbols for one gene. Afterwards, the list generated by the first approach was compared with the list of KEGG and the two groups were found to agree in approximately 65% of the genes. This is the case because either not all enzymes have IUBMB IDs, or the databases are not updated regularly, or because KEGG is not all-comprehensive

(121). After removing duplicates by keeping ensemble IDs are reference, 2427 unique enzymes were obtained representing a large fraction of characterized enzymes in human cells.

2.11 Clustering and Heatmaps

For all cluster analysis, Euclidean distance and Ward's Agglomerative Hierarchical Clustering Method (ward.D2) were used (122). This was not a random choice, but it was found to give the best separation among 2 or more groups, and it was computationally more feasible as compared to correlation-based distance calculation. All clustering analysis was carried out in R. Heatmaps were generated by using either *pheatmap* or *gplot* packages, respectively, and they are indicated in the relevant figures. All expression values of the genes were taken from log2-transformed for microarray and VST-transformed data tables for RNA sequencing and scaled by row before cluster and heatmap visualization. In microarray data, for the cases of genes with multiple probes, only the probe with the highest variance was retained while the others were discarded. This was done by calculating the variance of the entire dataset, re-order in descending order, and keeping only one unique gene symbol.

2.12 Survival Analysis and Gene Signature

For TCGA-LIHC survival analysis, clinical data downloaded from TCGA were used. For patients with more than one sequenced sample, only one was retained and those without follow-up survival data were removed. Survival period was converted into months on the basis of 365-days year. We used VST-data to avoid spurious correlations arising from adding constants before log2-transformation. For survival analysis, *surviver* and *survminer* packages were used and significance level was set as log-rank p-value < 0.05. A similar procedure was followed for microarray data used as external validation cohorts. 75% of patients in TCGA-LIHC cohort were

used to train a model starting with 17 candidate genes. Initially, *MyStepwise* package was used to reduce the number of genes into 4 which were used to create a cox regression model and fit it into two independent datasets; GSE76427 and GSE14520, but the absence of AKR1B15 probe in GSE14520 dataset prevented us from seeing the real power of this model to generalize. As a result, we used a second approach with *glmnet* package based on least absolute shrinkage and selection operator (LASSO) regularization (123) and obtained a signature based on 2 genes only. Further details are provided in the relevant section.

2.13 Data Manipulation and Other Visualization Graphs

All data manipulation such as cleaning, rearrangements and transformations were done with *dplyr* package or other necessary packages from *tidyverse*. All bar charts, pie charts, boxplots, radar plots, volcano plots, scatter and correlation plots were done using *ggplot2* package.

2.14 Software and Statistical Analysis

All the analysis for this study were carried out in R programming language on Linux operating system, Ubuntu 20.04 LTS distribution. Any other software is mentioned specifically within the text. All the R code for downloading and processing the data, as well as statistical analysis for this entire study is available in Github repository (https://github.com/ilirsheraj/Thesis).

CHAPTER 3

RESULTS AND DISCUSSION

This chapter consists of four main sections. The first section summarizes the results of a pan-cancer study on the expression of metabolic genes across 20 tumors and their matched adjacent tissues. Detailed analysis and results are available in a published manuscript (DOI:10.1038/s41598-021-93003-3). Sections 2-5 show detailed analysis of LIHC metabolism with an emphasis on the AKR1B10 and PPP pathway.

3.1 A Pan-Cancer Transcriptomic Study Showing Tumor-Specific Alterations in Cancer Metabolism

We aimed to evaluate transcriptome data to determine metabolic rewiring in 20 different cancer types with a specific focus on the expression of rate-limiting enzymes in different biochemical pathways. Differential expression analysis of genes involved in carbohydrate metabolism, including their transportation across plasma membrane, glycolysis, gluconeogenesis, PPP and hexosamine biosynthetic pathway; fatty acid metabolism-related processes of lipid biosynthesis and modification, beta oxidation, ketone body oxidation and biosynthesis; amino acid metabolism-related processes of amino acid transportation across plasma membrane, transamination reactions, amino acid oxidation, polyamine pathway, folic acid and urea cycles; and finally electron transport chain (ETC), TCA and anaplerotic reactions, were carried out.

For this purpose, RNA-Sequencing read counts of 1386 samples of tumors and their adjacent normal tissues constituting 20 different tumor cohorts were downloaded from TCGA using *TCGABiolink* package. Differential expression between tumor

and normal samples for each cohort was carried out with *DESeq2* package. The metabolic genes included in this study were mostly based in information available on Kyoto Encyclopedia of Genes and Genomes (KEGG) (https://www.genome.jp/kegg/pathway.html) and Molecular Signature Database (MsigDB) (https://www.gsea-msigdb.org/gsea/msigdb/index.jsp). The relevant genes were compared to data available in the literature; some genes that were deemed irrelevant for our purpose were removed, and others not available in any of the above collections were added manually.

3.1.1 Carbohydrate Transport, Glycolysis, PPP and Hexosamine Pathway

My analysis showed enhanced glycolytic flux in almost all tumors, starting from hexose transporters in plasma membrane. Many transporters, particularly GLUT1 which has the highest affinity for glucose, mannose and galactose (124) were upregulated to various degrees in almost all tumors. A similar pattern was observed for enzymes controlling the main steps of glycolysis: hexokinases (HKs), especially the high-affinity HK2 and HK3, phosphofructokinase (PFKs) and pyruvate kinases (PKs). I also showed the overexpression of lactate transporters (MCT1-4), in particular MCT-4 which is known for its importance in exporting lactate from highly glycolytic cells (125). All these enzymes have been shown to be essential in driving glycolytic flux in tumors (126). Different tumors were observed to show upregulation of different isoforms of the same enzymes and transporters. An upregulation of PPP was observed, which reinforces the importance of NADPH for both biosynthesis and oxidative stress mitigation, two essential processes for tumor survival and proliferation (127). Another important pathway that integrates carbohydrate, lipid, nucleotide and amino acid metabolism is the hexosamine biosynthetic pathway (128). While the pathway was generally upregulated in tumors, its degree of change was not as high as glycolysis and PPP. The number of genes up- and downregulated in each tumor are shown in the radar plot in Figure 3.1.

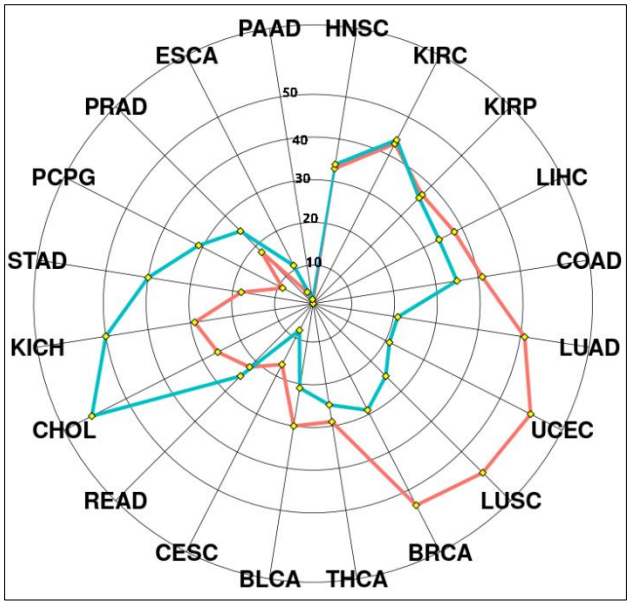

Figure 3.1 Carbohydrate Metabolism Differentially Expressed Genes.

Blue line represents downregulated and red line upregulated genes (FDR <0.05, LFC > 0.5). Numbers are shown within the plot.

Another important pathway that integrates carbohydrate, lipid, nucleotide and amino acid metabolism is the hexosamine biosynthetic pathway (128). While the pathway was generally upregulated in tumors, its degree of change was not as high as glycolysis and PPP. The number of genes up- and downregulated in each tumor are shown in the radar plot in Figure 3.1.

3.1.2 Fatty Acid Metabolism

Fatty acids (FAs) metabolism had been overlooked for years in cancer research but its importance has been recognized recently. FAs are important for many cellular processes. As they are the constituents of plasma and organelle membranes, they are involved in cellular signaling, and when oxidized produce immense amounts of energy per molecule as compared to carbohydrates (129–131). Overall, our analysis shows significant downregulation of genes involved in fatty acid oxidation (FAO) in different tumor types, both the main beta oxidation pathway and ketone body

breakdown. These findings support recent studies that have found that patients undergoing intermittent fasting or kept under ketogenic diet showed better response to cancer chemotherapy (132).

This reduction could be for two reasons: first, FA oxidation requires oxygen as it takes place in mitochondria while tumors grow under hypoxic conditions, and since the general state of tumors is proliferation, synthesizing and breaking FAs down simultaneously would be a futile cycle. The expression of enzymes involved in fatty acid biosynthesis (FAS) on the other hand was found to be significantly upregulated in most of the tumors. In addition, we also observed major changes in FA modifying enzymes which can lead to significant changes in tumors' overall lipidome, affecting their membrane composition, cellular signaling, growth and proliferation. Figure 3.2 summarizes the number of FA metabolism genes that are up- and downregulated in 20 tumor cohorts.

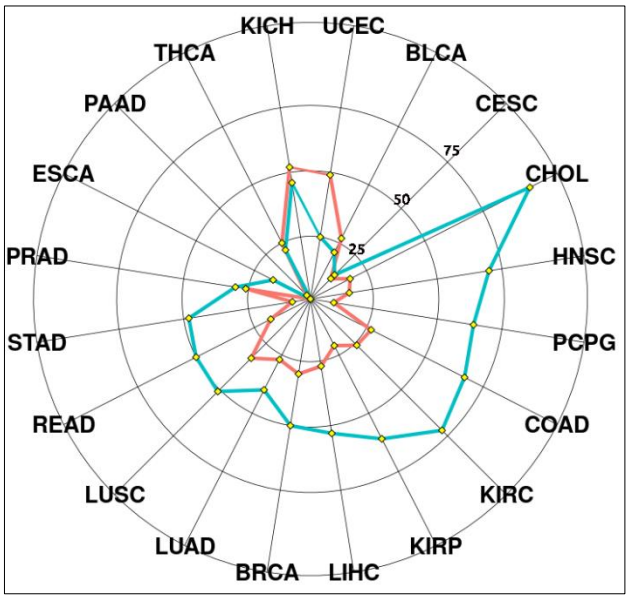

Figure 3.2 Fatty Acid Metabolism Differentially Expressed Genes.

Blue line represents downregulated and red line upregulated genes (FDR <0.05, LFC > 0.5). Numbers are shown within the plot.

3.1.3 Amino Acid Metabolism

Amino acids (AAs) are essential for cell survival and proliferation because they are the basic building blocks of proteins, are essential for oxidative stress response especially through glutathione and proline, and they are utilized for energy through anaplerotic and trans-amination reactions. AA availability in the cell is associated with an overall anabolic state while their deficiency is associated with overall catabolism (133–135). Eukaryotic cells cannot synthesize most of the AAs so they have to be obtained from the diet (136).

We observed the overexpression of AA transporters in the plasma membrane, especially Glutamine, Serine, Cysteine and Glycine among others, showing increased AA flux into tumor cells. AA oxidation on the other hand was generally decreased, indicating that tumors, despite increasing AA flux (auxotrophs), do not utilize them for energy but preserve them either for coping with oxidative stress or for protein synthesis. This idea is supported by increased expression of enzymes involved in Serine, Proline, and Arginine biosynthesis, as well as downregulation of urea cycle enzymes, which collect amino groups produced by deamination reactions during AA breakdown (137). In addition, we found significant overexpression of genes involved in the polyamine pathway which has recently been found to be important for cancer progression and aggressiveness (32). Figure 3.3 summarizes the number of differentially expressed genes involved in AA metabolism and transport.

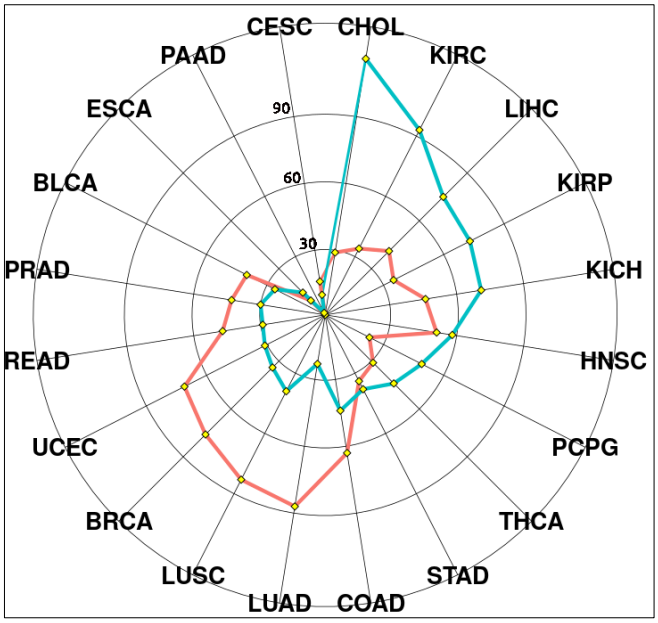

Figure 3.3 Amino Acid Metabolism Differentially Expressed Genes.

Blue line represents downregulated and red line upregulated genes (FDR <0.05, LFC > 0.5). Numbers are shown within the plot.

3.1.4 TCA, Anaplerosis and ETC

TCA, besides being the main pathway for ATP synthesis, also functions as the integrative point of many biological pathways such as carbohydrates, AA and FA oxidation and biosynthesis. Thus, controlling its flow is essential for tumor growth and proliferation (138). My data shows that the changes in anaplerosis and TCA are mostly tumor-specific. However, we observed a general decrease in the expression of pyruvate dehydrogenase and pyruvate transporters (MCP1/2) in the mitochondria. Depletion of MCP1/2 has been shown to force cells to shift their energy sources to glutaminolysis in normoxic conditions (139). The expression of IDH1, IDH2 and ME1 also varied among different tumors. I also found that in most tumors, the expression of mitochondrially-encoded, core enzymes of the ETC were significantly downregulated while those with higher ETC enzyme expression had increased expression of uncoupling enzymes, a group of mitochondrial channels that uncouple

the electron gradient, turning it into heat rather than ATP (140–142). Finally, we also found a general downregulation of electron transfer flavoproteins (ETFs) which transfer electrons from the initial step of beta oxidation or branched AA oxidation in the mitochondrial matrix to complex II of ETC. Also, changing between tumors, either some or all proteins of the glycerophosphate shuttle, an important electron transfer system that oxidizes cytosolic NADH and transfers its electrons to ETC were also downregulated (Figure 3.4).

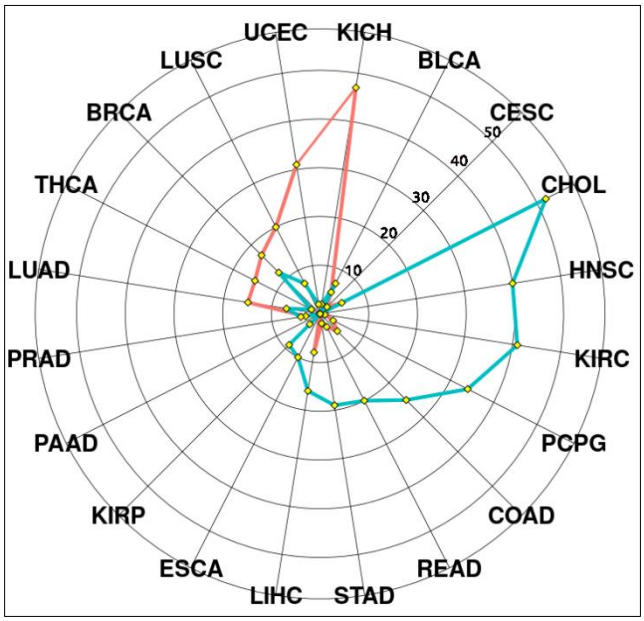

Figure 3.4 TCA, Anaplerosis and ETC Metabolism Differentially Expressed Genes. Blue line represents downregulated and red line upregulated genes (FDR <0.05, LFC > 0.5). Numbers are shown within the plot.

Overall, the expression of carbohydrate and AA transporters was found to be significantly overexpressed in tumors. In addition, tumors had higher expression of glycolytic, PPP and hexosamine enzymes. Essential AAs were not utilized for energy generation but were funneled into protein synthesis while amino groups were used for the synthesis of other AAs through transamination and polyamines. Tumors also showed decreased FA and ketone body oxidation, a surprising state since these

molecules are very rich energy sources. FA biosynthesis on the other hand was increased and the expression of FA modifying enzymes showed major changes. Overall, the decrease in FA oxidation can be explained by the general lack of oxygen and an overall anabolic state of the cell, while dysregulation of FA modifying enzymes needs careful consideration since the lipidome profile has major implication in cellular physiology (143). Contrary to data generated from cell culture experiments, I did not find very significant changes in anaplerotic reactions and TCA. This is likely because tumor samples grow and proliferate in a dynamic environment. The overall findings of this study are shown schematically in Figure 3.5.

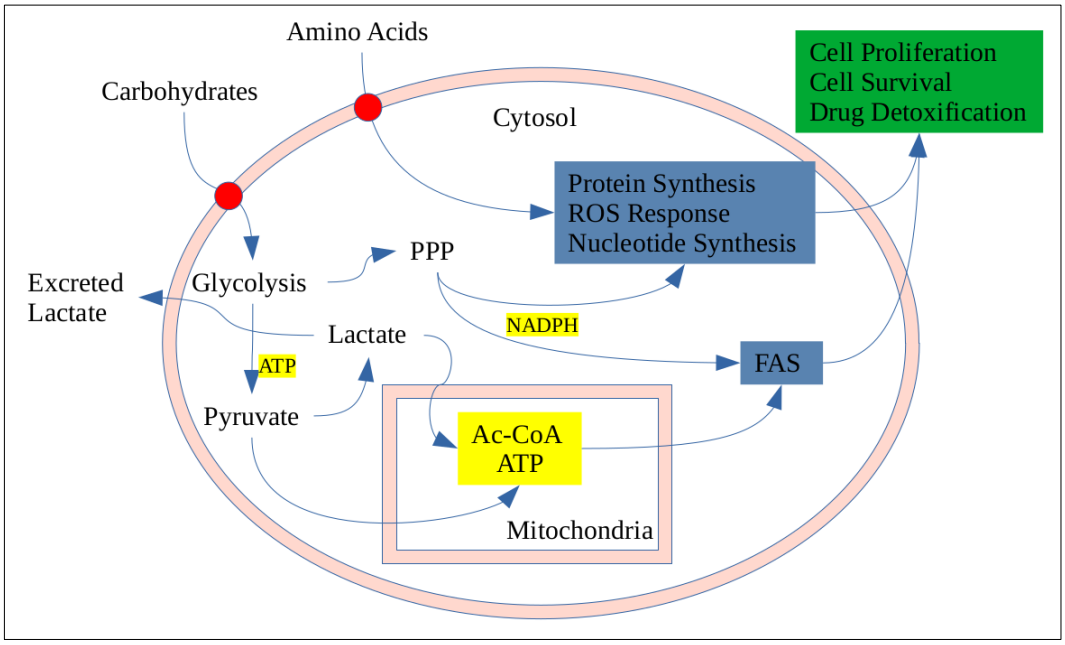

Figure 3.5 Schematic Drawing Showing the Flow of Metabolites in Tumors.

3.1.5 Expression NADPH Metabolizing Enzyme

Figure 3.5 indicates that NADPH is one of the central players in metabolic deregulation in cancer cells. The major NADPH synthesizing pathways in the cell are the PPP, one-carbon cycle (OCC) and some enzymes involved in anaplerotic

pathways (33). Among the NADPH utilizing enzymes is AKR1B10, which requires electrons from NADPH to carry out its enzymatic reactions. Therefore, we evaluated the expression of AKR1B10 and several NADPH-synthesizing enzymes shown in Figure 1.3 across all 20 tumor cohorts and represented them in a heatmap (Figure 3.6).

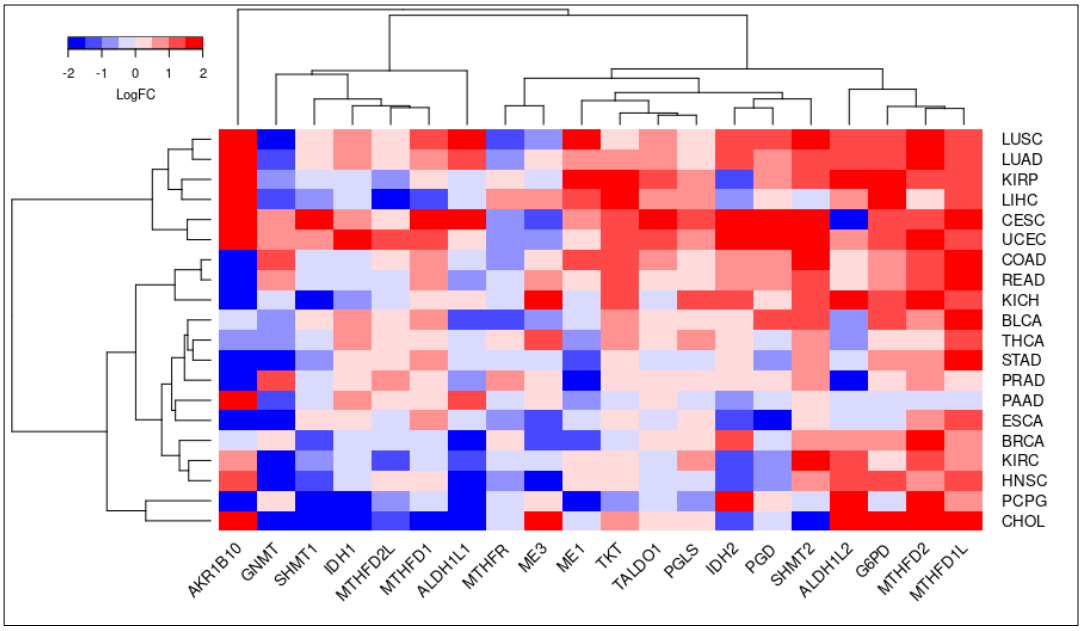

Figure 3.6 Expression of AKR1B10 and NADPH Synthesizing Enzymes Across TCGA Cohorts.

Heatmap was constructed with Euclidean distance and Ward.D2 linkage. Color bar at the top shows expression changes in LFC.

AKR1B10 was significantly upregulated in 8 tumors while G6PD was significantly upregulated in more than half of the tumors. Of relevance to the current study, AKR1B10 together with genes involved in especially in PPP but to various extent also in OCC and anaplerosis were highly upregulated in LIHC tumors as compared to adjacent normal tissue. These data support the building of our hypothesis for a common regulatory network for enzymes utilizing NADP(H) as a coenzyme in LIHC.

3.2 Stepwise LIHC Development

3.2.1 Microarray Gene Expression

LIHC is a very heterogeneous tumor and its progression generally involves multiple stages. In order to understand the stage-wise progression at the molecular level, raw gene expression data from the dataset GSE6764 as well as the relevant clinical information were download from the Gene Expression Omnibus (GEO). In this study, the authors classified 75 liver samples from 48 patients into different categories and stages according to conventional pathological classification appropriate for the liver (Table 3.1). Afterwards, overall gene expression was evaluated using microarray technology (105).

Table 3.1 Sample Pathology

Sample Type	*Number*
Healthy Tissue	10
Cirrhosis Viral	10
Cirrhosis Non-Viral	3
Low-Grade Dysplasia	10
High-Grade Dysplasia	7
Very-Early HCC	8
Early HCC	10
Advanced HCC	7
Very-Advanced HCC	10

It should be emphasized that healthy liver tissues were not taken from livers affected either by cirrhosis or tumor, but from healthy patients without any liver pathology. Cirrhosis samples were also of two types, infected with HCV and non-viral. There were two grades of dysplasia, low-grade and advanced. LIHC samples were separated into very-early and early LIHC with well to moderate differentiation, and

advanced and very advanced LIHC stages with poor differentiation and macrovascular invasion (105).

Using data from GSE6764, we could also follow the evolution in expression of all metabolic genes represented in the microarray. More importantly, this dataset made it possible to check whether AKR1B10 and PPP genes' expression changes in very early pathological stages of liver, or it changes in more advanced stages of tumor. After raw data were downloaded, all samples were normalized together using robust multiarray averaging (RMA) method as described in Chapter 2.

Next, an unsupervised cluster analysis for the whole transcriptome was performed to see how well the transcriptome profile agreed with pathological classification. The best separation was achieved by using Euclidean distance and Ward.D2-based linkage (Figure 3.7).

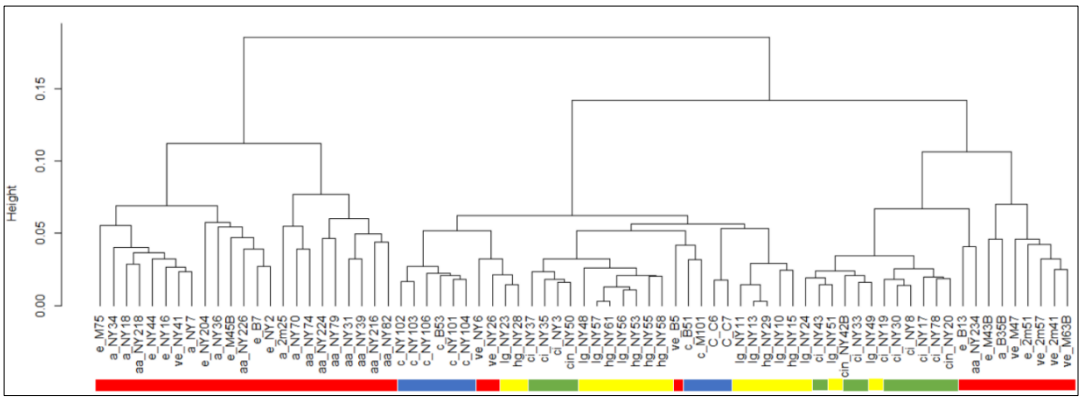

Figure 3.7 Unsupervised Hierarchical Clustering of GSE6764 Transcriptome.
Color labeling: red represents all tumor samples from very early to very advanced. Blue represents normal samples, green cirrhotic and yellow dysplastic. More detailed labeling is given by the initial letter codes as follows: "c" stands for control, "ci", cirrhotic, "cin", cirrhotic non-viral, "lg", low-grade dysplasia, "hg", high-grade dysplasia, "ve", very early, "e", early, "a", advanced and "aa" very advanced.

Clustering analysis shows that (with few exceptions) tumor samples were well-separated from the others, while normal, cirrhotic and dysplastic samples clustered closer to each other. This is also in agreement with the cluster analysis in the original article (105). Since there are few samples for each stage, we decided to aggregate them into larger but more pathologically heterogeneous groups to increase statistical power, an approach also used in the original study (105). However, before this step, we checked the number of significantly differentially expressed genes between the substages. For this purpose, differential gene expression (DGE) analysis between viral cirrhosis versus non-viral cirrhosis, low- versus high-grade dysplasia, very-early versus early LIHC and advanced versus very-advanced LIHC were carried out using the *limma* package. In all cases, no significantly differentially expressed genes (DEGs) were found (FDR < 0.05) as shown in Figure 3.8. Although this may introduce some loss of data because of sample heterogeneity, the increase in statistical power from more samples overall outweighed the drawbacks.

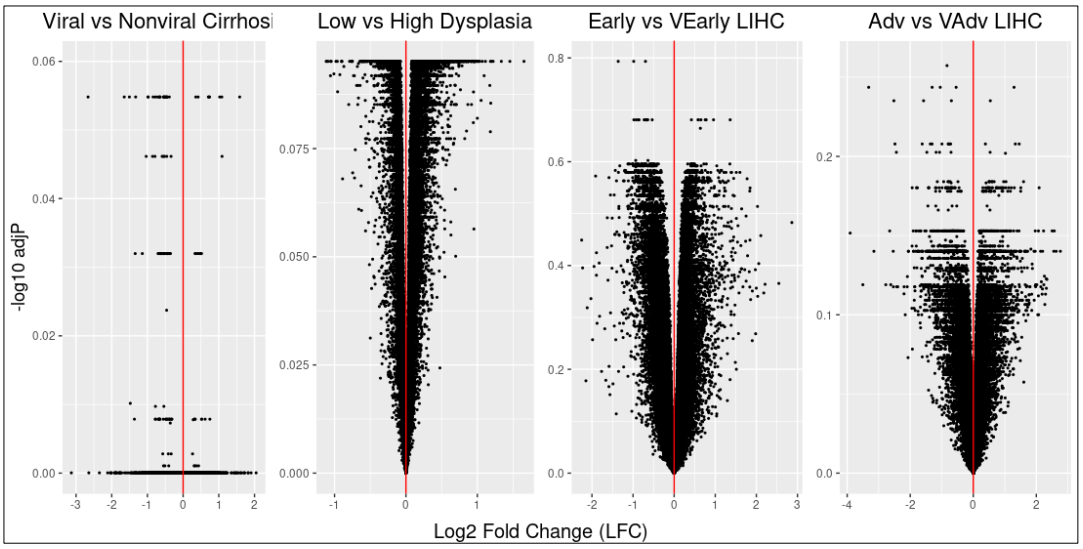

Figure 3.8 Volcano Plots Showing DGE Between Different Substages in GSE6764. Contrasts are shown on top of each plot, y-axis shows negative log10 of adjusted p-value, and the x-axis the log2-fold change. The horizontal line separated the two conditions under investigation. None of the genes reached statistical significance (FDR < 0.05).

3.2.2 Differential Expression Analysis Between Aggregated Stages

Since no gene was found to show a strong differential expression, we aggregated the samples into 5 groups and renamed them as shown in Table 3.2.

Table 3.2 Aggregated Sample Pathology

Sample Type	*Number*
Healthy Tissue	10
Cirrhosis Viral	13
Dysplasia	17
Early HCC	18
Advanced HCC	17

We next carried out DGE analysis between each pathological stage and normal healthy liver samples using *limma*. Results show an increasing number of DEGs with advancing disease stage as expected (Figure 3.9).

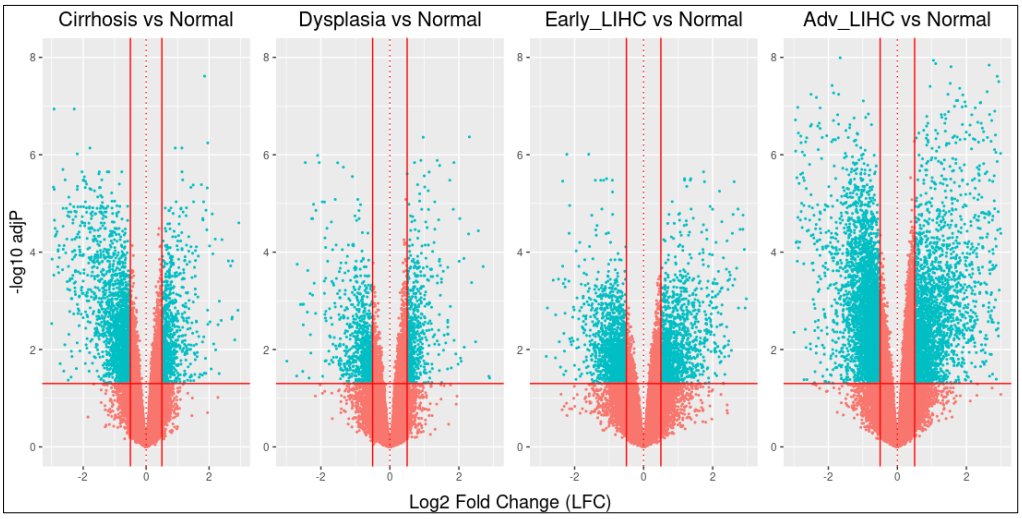

Figure 3.9 Volcano Plots of DGE Between Normal and Pathological Tissue.
Vertical lines show the 0.5 LFC and horizontal line the FDR = 0.05, the cut-offs used for significance. All genes passing these filters are colored in blue.

Using FDR<0.05 and LFC >0.5, we carried out gene ontology (GO) enrichment analysis using *clusterProfiler* (120) package with FDR <0.01 for genes up- and downregulated in all pathological stages each compared to the normal tissues. There were multiple GO terms that showed enrichment for all comparisons, with majority of them being redundant because the same gene is found in more than one GO term. In order to make the results simpler, similar GO terms were removed by first using *simplify* function of the same package and then manually choosing the most representative ones. Figure 3.10 shows the top GO terms enriched from differentially expressed genes in cirrhosis when compared to normal samples.

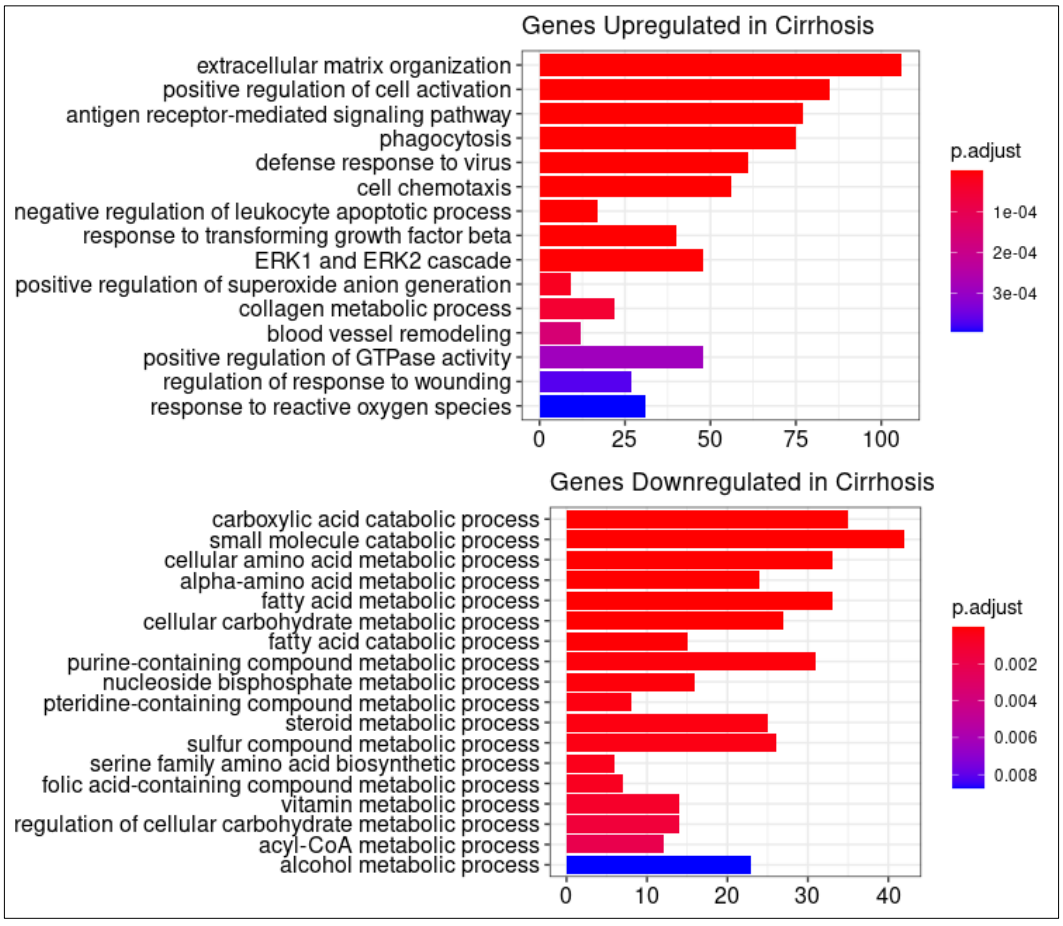

Figure 3.10 GO Term Enrichment for Genes Up- and Downregulated in Cirrhosis. The numbers at the bottom of each graph show the number of genes per each GO term.

As the figure shows, many metabolic processes were downregulated in cirrhotic cells such as amino acid and fatty acid catabolism. Of major importance could be the downregulation of folate cycle and vitamin metabolism, but the real biological significance of these processes, to our knowledge, has not been studied. On the other hand, cirrhosis showed increased expression of genes constituting multiple GO terms related to inflammation, wound healing, viral defense and extracellular matrix organization. All of these pathways are expected to be enriched in the state of cirrhosis. Figure 3.11 shows the top GO terms enriched from differentially expressed genes in dysplasia as compared to normal. Here also we can see an enrichment of GO terms relevant to inflammation and extracellular processes. On the other hand, genes downregulated in dysplasia (highly expressed in normal tissues) are again of a metabolic nature.

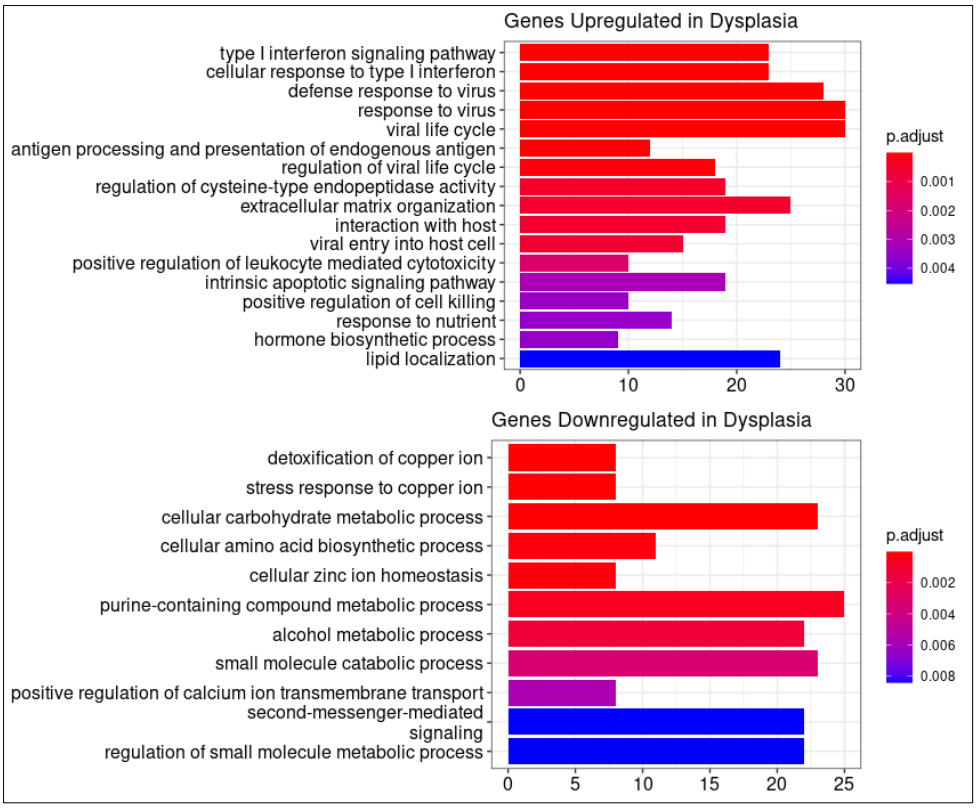

Figure 3.11 GO Term Enrichment for Genes Up- and Downregulated in Dysplasia. The numbers at the bottom of each graph show the number of genes per each GO Term.

GO term enrichment of genes upregulated in early LIHC stages as compared to normal tissues showed an upregulation of biological processes associated with chromosomal segregation, cell cycle checkpoint control and cellular replication, all of which are typical tumor-related processes. In contrast, genes downregulated (higher expression in normal tissues) were related to metabolic processes, especially lipid catabolic pathways (Figure 3.12). These results are in agreement with pan-cancer analysis shown in section 3.1 of this chapter, and this could be significant as liver is known to preferentially oxidize fatty acids under normal conditions (144).

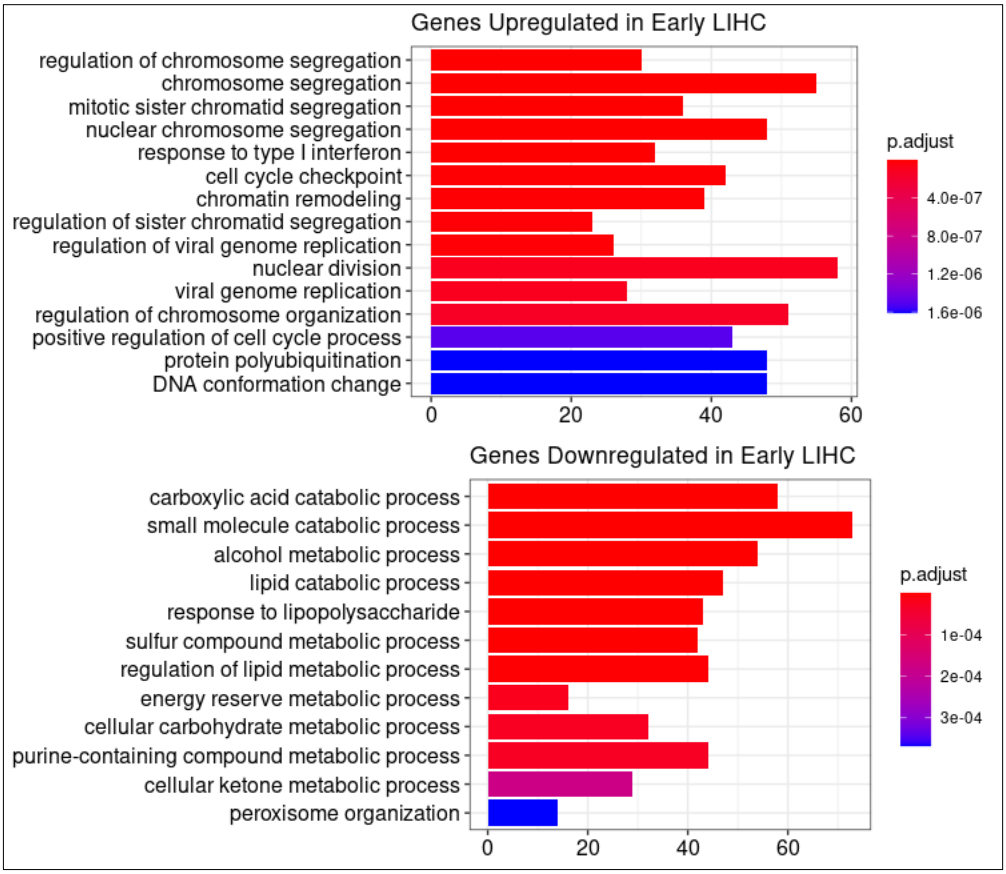

Figure 3.12 GO Term Enrichment for Genes Up- and Downregulated in Early LIHC. The numbers at the bottom of each graph show the number of genes for each GO Term.

Finally, a similar analysis between advanced LIHC and normal tissues (Figure 3.13) showed a similar picture to that of early LIHC, but the number of DEGs was much higher and their LFC was larger, as would be expected from more advanced tumors.

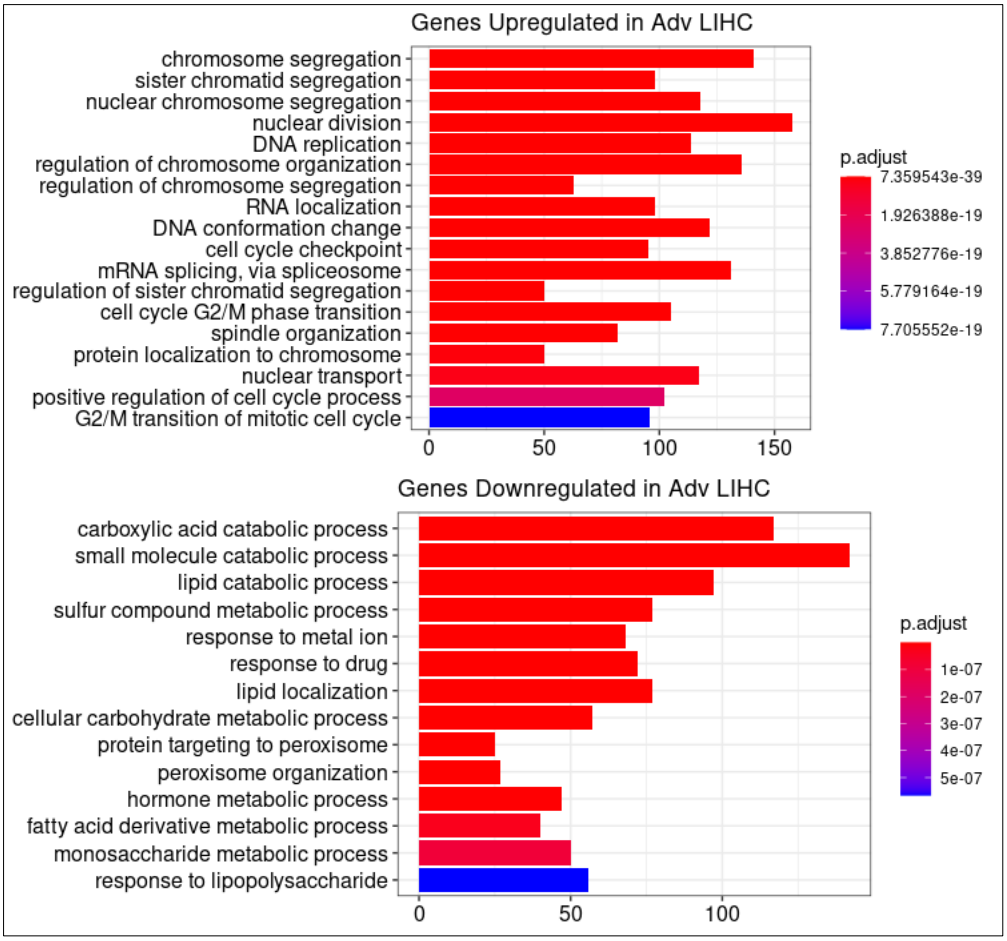

Figure 3.13 GO Term Enrichment for Genes Up- and Downregulated in Advanced LIHC.

The numbers at the bottom of each graph show the number of genes for each GO Term.

While the upregulation of genes involved in cell cycle, DNA replication, RNA metabolism and chromosomal maintenance in tumor was not a surprise, the downregulation of many genes involved in metabolism is important because of the normal functions of liver as an organ. Liver is the heart of metabolism and based on

the available nutrients in the blood, it manages their distribution as needed. Liver is known to use carbohydrates only when they are available in high very amounts in the blood by regulating their entry both into hepatocytes and entry into glycolysis by hexokinase IV which is known for its low affinity. The data from our pan-cancer study showed that this condition was reversed in LIHC. In addition, the liver preferentially oxidizes fats for both its own energy and for converting their end products into ketone bodies, which are released into the blood to be metabolized by other tissues (144). The liver also is the main organ where amino acids are converted into glucose by a process that operates in opposite direction of glycolysis called gluconeogenesis to maintain optimal sugar levels in the blood. The gluconeogenic enzymes were found to be downregulated in LIHC. Therefore, based on the results we obtained from exploratory analysis in this dataset as well as others, we decided to evaluate the expression of a comprehensive number of metabolic enzymes across all stages in this dataset and identify enzymes that were consistently up- and/or downregulated throughout the whole evolutionary stages of LIHC.

3.2.3 Metabolic Gene Expression Among LIHC Stages

Since the primary aim of this study is to evaluate the changes in metabolism-related genes of LIHC, the next step was to evaluate the expression of metabolic enzymes across all different pathological conditions found in the GSE6764 dataset. Using the approach described in detail in section 2.10, 2427 unique enzymes were obtained. The microarray probes were converted into ensemble IDs by using Affymetrix annotation package *hgu133plus2.db* and the probes pertaining to the enzymes filtered out. Eventually, 5220 probes representing 2271 unique enzymes were obtained.

To reduce the noise introduced by multiple probes representing a single gene, the variance of each probe was calculated across all samples. For each enzyme, the probe with the highest variance was retained since larger variance carries more relevant information about a gene, while taking probe average would introduce a lot of noise

into the data. This approach may introduce some bias when thousands of genes are considered, but our main purpose at this stage was to visualize and cluster the samples as best as we could. Finally, the expression values of each gene were scaled and used to generate a heatmap whose distance and linkage were calculated by various methods. However, Euclidean distance and Ward's linkage yielded the best separation shown by the heatmap (Figure 3.14).

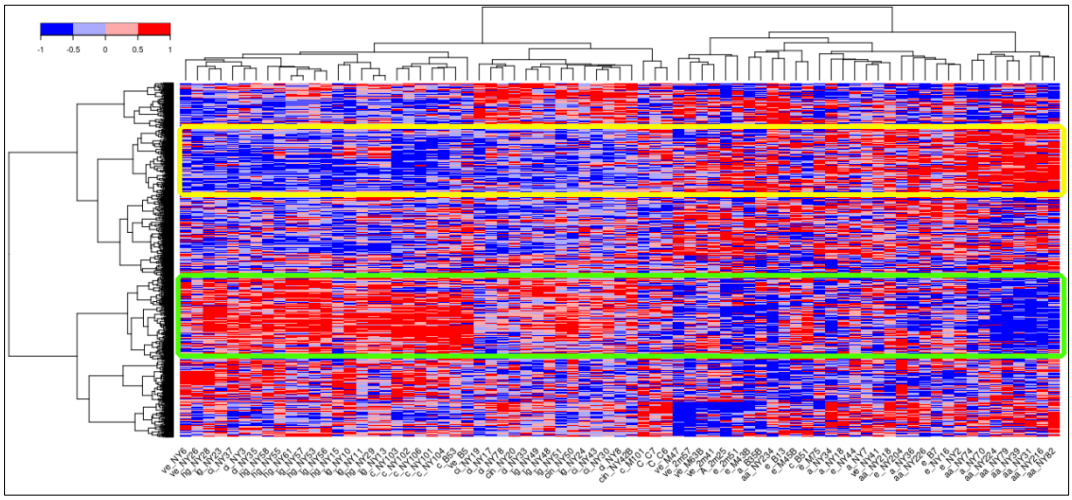

Figure 3.14 Enzyme Expression Heatmap for GSE6764.

The heatmap shows the expression of 2271 enzymes. Cluster 1 and 2 which constitute the most distinct groups in tumor and other groups are marked by yellow and green color boxes.

We observed that no matter what the tumor's pathological stage was, metabolic enzymes could be used to separate them from normal, cirrhotic or dysplastic samples, but the same genes did not contribute towards the separation of cirrhosis and dysplasia from normal healthy samples. This is in agreement with GO term enrichments shown in Figures 3.10-3.13. In the heatmap we can see two well-separated clusters for tumor samples and the rest. The constituent genes were retrieved and used for GO term enrichment (Figure 3.15).

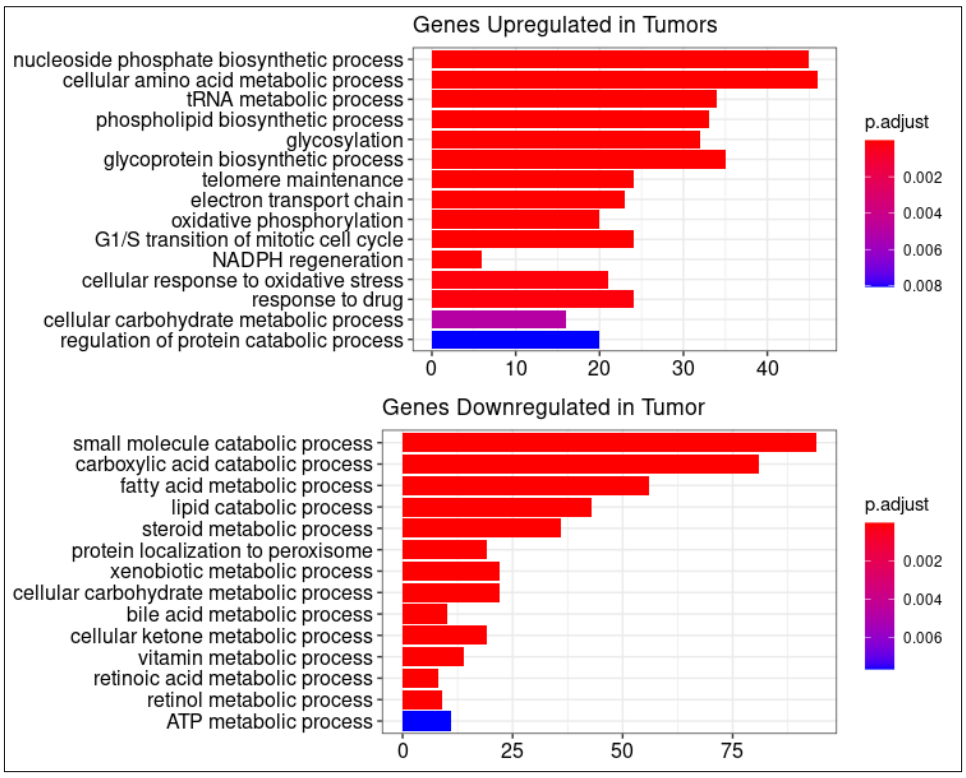

Figure 3.15 GO Term Enrichment for Cluster 1 and Cluster 2 from Heatmap.

The most representative GO terms were filtered out of hundreds of GO terms as it is explained in detail in chapter 2 and above.

In full agreement with previous analyses, the first cluster which represents genes upregulated in tumors with respect to normal, cirrhosis and dysplasia shows GO terms involved in cancer proliferation such as tRNA loading, nucleotide biosynthesis, cellular glycosylation, oxidative stress, NADPH biosynthesis, lipid biosynthesis and modification etc. The AKR family of enzymes AKR1B10 and AKR1C1/2/3 were also present in this cluster together with PPP genes as well as other oxidative-stress related enzymes such as GPXs and TXNDs. In contrast, the second cluster is made of genes involved in lipid and ketone body oxidation, amino acid metabolism, alcohol, vitamin and retinol metabolism among others. The expression of all these genes is decreased in LIHC, more so in advanced stages (Figure 3.14). These results are also in full agreement with our RNA-Seq analysis, showing that liver has major rewiring of its metabolism to maximize growth and

proliferation. Overexpression of many drug-metabolizing enzymes as well as PPP which is a major contributor of NADPH in the cell could be very important factors in making chemotherapeutic drugs ineffective, thus these pathways require more attention in future research.

3.2.4 Expression of AKR1B10 and genes of PPP

Since AKR1B10 and enzymes of the PPP were found to be clustered in the highly differentially expressed subset of genes and significant enrichment of the GO term "NAPDH regeneration" was observed from the list of metabolically important enzymes, we evaluated these two pathways further. The expression of AKR1B10 is already known to have a high potential for a biomarker for the early detection of liver cancer. We therefore used data from GSE6764 to evaluate the expression of AKR1B10 and PPP genes in the normal tissue and follow their evolution through various pathological stages leading to advanced LIHC (Figure 3.16).

As shown in Figure 3.16, the expression of AKR1B10 increased in a stage-wise manner leading to LIHC. However, in agreement with the available literature, not all tumors show consistent elevation in AKR1B10 expression as compared to normal, as reflected in the large variation observed in the different stages of the disease. The increase in the expression of the PPP genes was more uniform in tumor samples, especially G6PD and TKT which started to increase in expression uniformly from very early stage, while no change between normal and cirrhotic/dysplastic samples was observed. On the other hand, PGD and TALDO1 showed moderate increase in tumors but expression values were very noisy with high variation within the groups.

When the aggregated samples were evaluated, the expression of AKR1B10 was seen to increase as the pathological grade of samples advanced from normal to LIHC (Figure 3.17). The expression of the PPP enzymes G6PD and TKT increased only in tumor, especially the advanced stages.

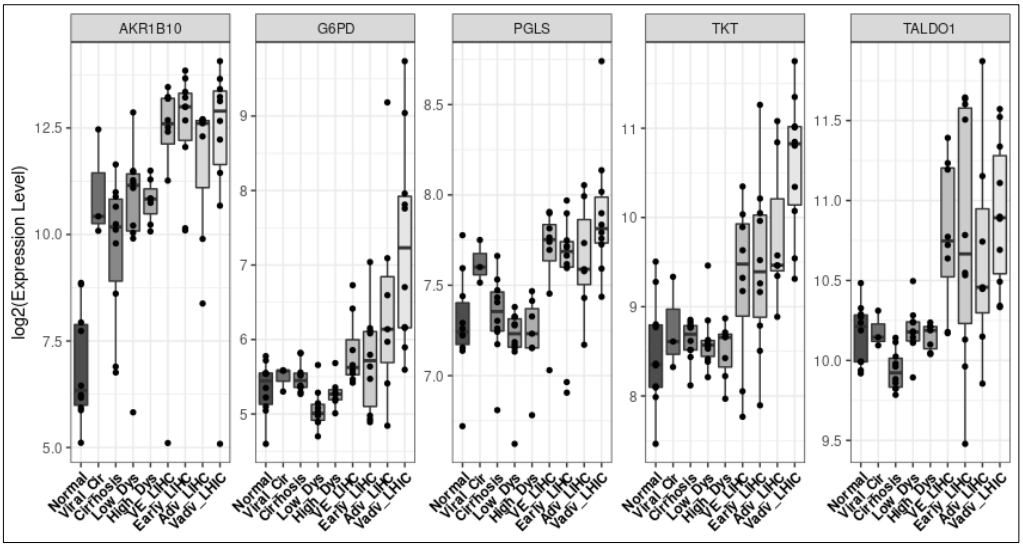

Figure 3.16 AKR1B10 and PPP Gene Expression Across Multiple LIHC Stages.

Gene expression is shown for every stage starting from normal, viral cirrhosis (Viral_Cir), cirrhosis, low-grade dysplasia (Low_Dys), high-grade dysplasia (High_Dys), very-early (VE_LIHC), early (Ear_LIHC), advanced (Adv_LIHC) and very advanced (Vadv_LIHC) liver tumor. Note that y-axis is not to scale for the purpose of clarity. Except PGD, all had statistical significance which is not shown for clarity purposes (One-Way Anova, $p < 0.05$).

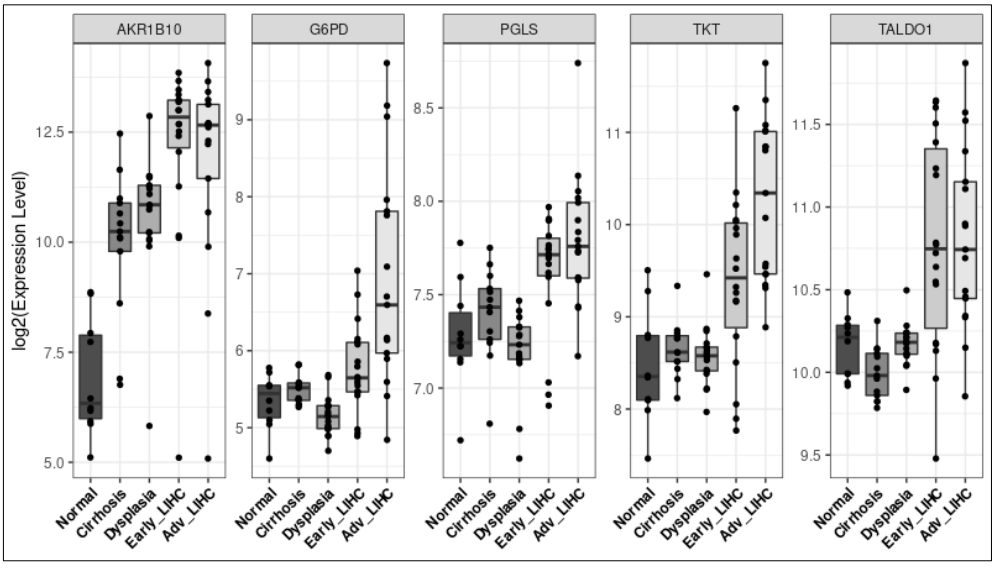

Figure 3.17 Expression of AKR1B10 and PPP in Aggregated Pathological Stages.

All except PGD showed statistical significance (One-Way Anova, $p < 0.05$).

Using the GSE6764 microarray dataset we established that the expression of AKR1B10 was upregulated at the very early stages of cirrhosis and dysplasia, followed by another increase in expression in LIHC from a very early stage. However, as has been shown in the literature, overexpression does not occur in all patients.

In addition, we also showed that the enzymes of the PPP were also upregulated in a stage-wise manner in LIHC but not in cirrhosis or dysplasia. Together with this pathway, a number of enzymes whose activity depends on NADPH such as those involved in oxidative stress, xenobiotic detoxification (AKR1C), lipid, cholesterol and nucleotide biosynthesis were all upregulated. On the other hand, with this dataset we reconfirmed results obtained by RNA sequencing in our pan-cancer study. Our data suggest that LIHC is characterized by decreased lipid and amino acid catabolism, decreased alcohol/aldehyde pathways, as well as dysregulations in bile acid pathway, all of them essential processes for a healthy liver (145).

3.3 TCGA RNA-sequencing of LIHC Normal and Matched Tumor Samples

To further support the data on the expression of AKR1B10 and PPP genes from GSE6764 dataset, we evaluated their expression in the TCGA-LIHC dataset which is based on RNA Sequencing, thus providing absolute expression values for any gene of interest. In addition, having normal matched samples to compare with the cancers is important to understand the real changes happening in tumors because multiple factors such as genetic make-up, lifestyle, race, sex and scores of other known and unknown variables potentially affecting the results drastically (146). This dataset makes it possible to remove all these variables which may have introduced a lot of noise in GSE6764 dataset. The LIHC dataset in TCGA is also more comprehensive since it contains data on patient survival, their whole genome sequence as well as

whole methylome. All these pieces of information will be integrated in order to get a clearer picture about the role and state of AKR1B10 and PPP genes in LIHC.

3.3.1 Differential Gene Expression of Normal and Matched Tumors

Raw read counts for tumor and their adjacent normal tissues for 50 LIHC pateints were downloaded from TCGA server using *TCGABiolink* (109) package as explained in chapter 2. Afterwards, matched samples were normalized and DGE analysis was carried out by *DESeq2* (111) using a general linear model (GLM) with single variable about samples' status (normal and tumor). First, simple exploratory analysis showed good separation of samples at the level of whole transcriptome with the exception of two tumor samples that were clustered with normals (Figure 3.18).

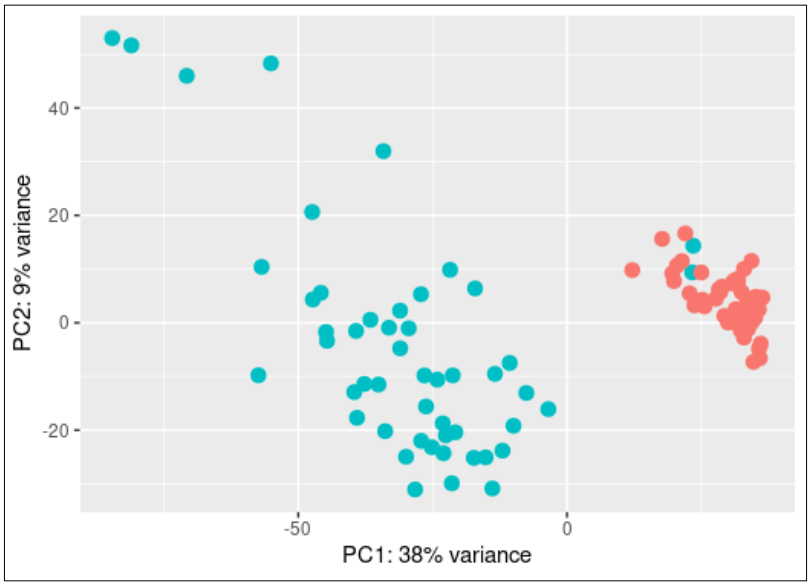

Figure 3.18 PCA Plot of Tumor and Matched Normal Samples.
The plot represents the whole transcriptome. Blue dots show tumor and red dots normal samples.

Figure 3.18 also confirms the well-known heterogeneity of LIHC. DGE analysis showed over 2000 and 5000 transcripts with significant differential expression

between normal and tumor samples (FDR < 0.05, 1LFC), respectively, the majority of them protein-coding genes but with a significant fraction of other genetic elements such miRNA, lincRNA and pseudogenes. A summary of the nature of those transcripts is shown in Figure 3.19.

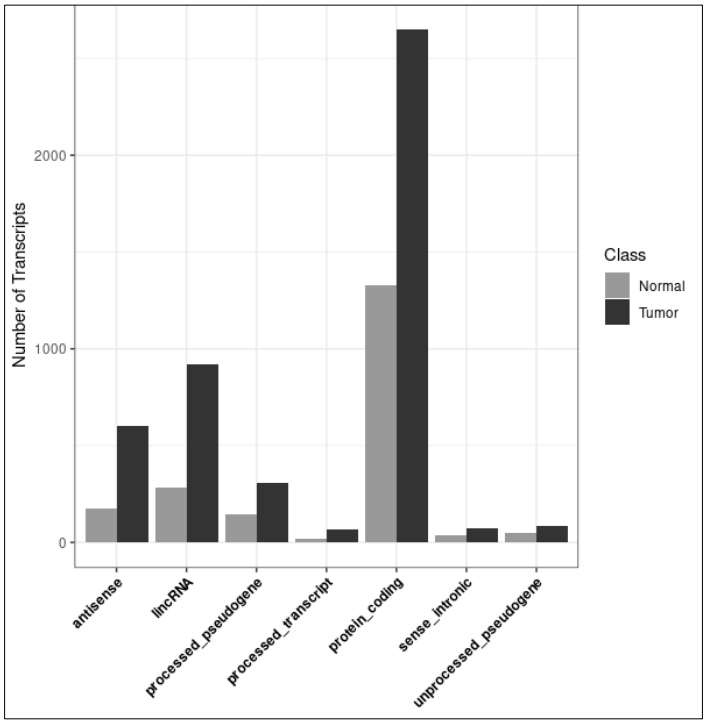

Figure 3.19 Summary of Significantly Differentially Expressed Transcripts Between Normal and Tumor Samples.

Next, we evaluated the expression of AKR1B10 and PPP genes. With the exception of PGD, all of the genes were significantly upregulated in tumors as compared to normals. Figure 3.20 also shows that the variance in the expression of AKR1B10 was very high in both normal and tumor, and the gene was upregulated in 80% of tumor samples as compared to normal and downregulated in 20% others. On the other hand, G6PD, TKT and TALDO1 had lower variation in expression. In addition, G6PD was upregulated in 85% of samples, while TKT and TALDO1 only in about 50%.

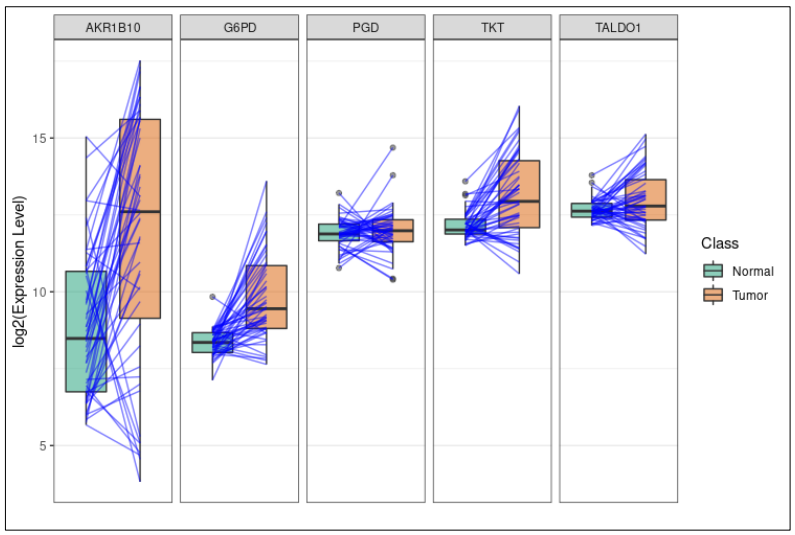

Figure 3.20 AKR1B10 and PPP Gene Expression in Tumor and Normal Samples.
The horizontal lines connect each tumor sample to its matched normal. All except PGD show highly significant change (Wald's test, $p < 0.001$).

These data confirmed the results obtained from the microarray dataset GSE6764, and also showed the change of these particular genes within patients. A representative sample of 50 patients has enough power to generalize these results over the population albeit with caution. We next evaluated the dependence of the expression of these genes for different grades (Grade and TNM), Neoplasm histologic state (G-Stage) and patient sex (Figure 3.21) as well as their correlation with Buffa (147), Ragnum (148) and Winter (149) hypoxia scores (Figure 3.22). The results in Figure 3.21 show a tendency of the expression of AKR1B10 to increase with increasing G-Stage and there was a significant change in expression between G1 and G3. On the other hand, G6PD showed a stage-wise increase in expression, a pattern observed in TKT, albeit not as consistent. The other two enzymes showed no significant change in expression. For tumor grade (I-IV), only G6PD showed an increase in expression as the tumor advanced, a pattern seen in the GSE6764 dataset while the other enzymes had no significant change. The expression of none of the enzymes across T-stage reached statistical significance except G6PD. Finally, all the enzymes had significantly increased expression in males compared to females.

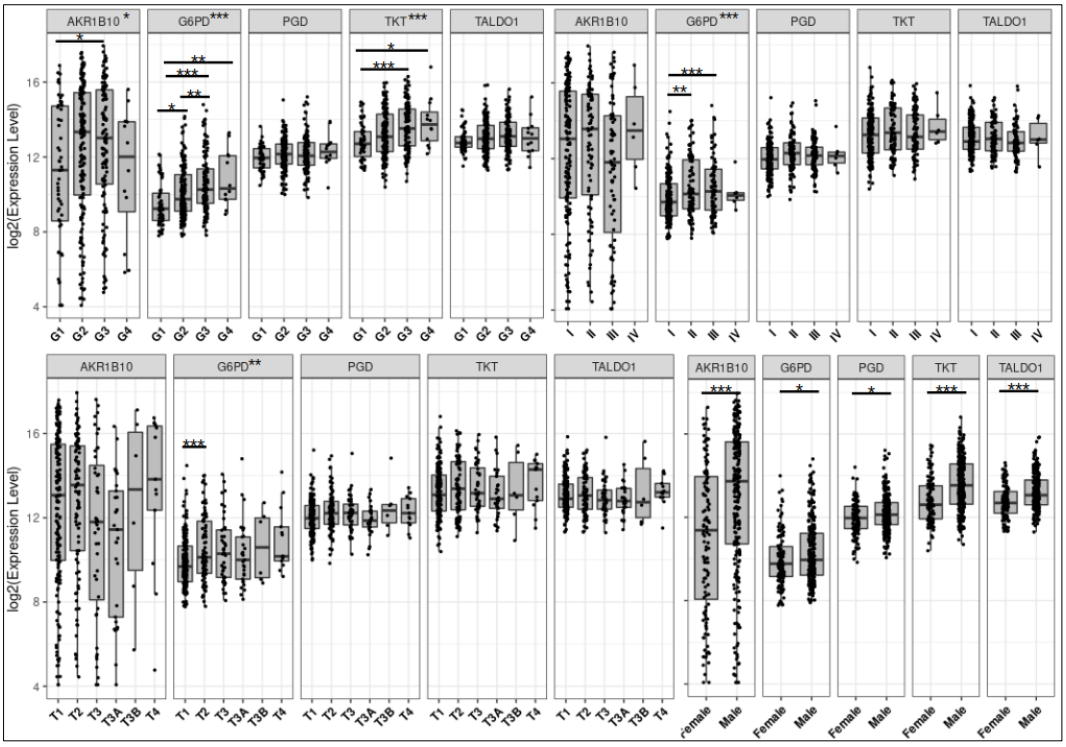

Figure 3.21 AKR1B10 and PPP Enzyme Expression Dependency.

Expression differences of the enzymes were checked with respect to Neoplasm histologic state (G-Stage), tumor stage (I-IV), T-stage (T1-T4) and gender (female versus male). Statistical significance was calculated by One-Way Anova with Tukey's post-hoc. Only bars of significant changes are shown for clarity. (*: $p < 0.05$, **: $p < 0.01$, ***: $p<0.001$).

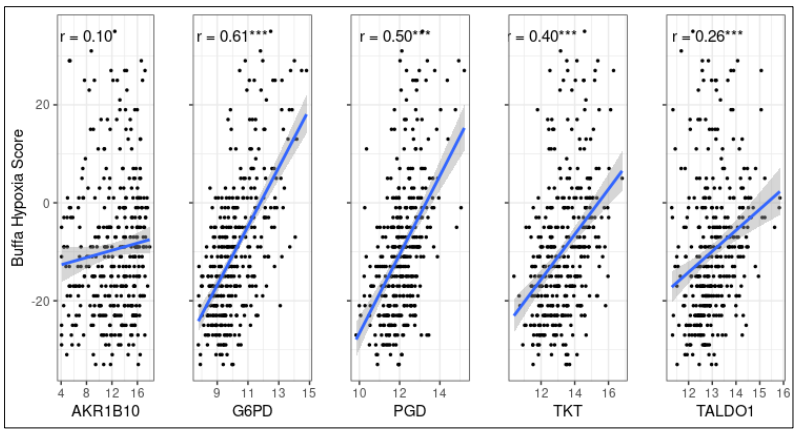

Figure 3.22 Correlations of Gene Expression with Buffa Hypoxia Score.

Pearson coefficients (r) and their significance are shown on top of scatterplots.

Correlation analysis of expression of our genes of interest with the 3 available hypoxia scores showed no correlation between AKR1B10 and any of the scores. G6PD on the other hand was strongly positively correlated with all the scores (Pearson r = 0.58-0.61, $p < 0.001$). PGD and TKT were also correlated with all scores albeit weakly, while TALDO1 showed no correlation with Ragnum Score (r = -0.02, p=0.7). Scatter plots showing the correlation of each of these genes with Buffa score are shown in Figure 3.22 while the plots of two other scores are omitted. These results indicate that PPP, especially its rate-limiting enzyme G6PD is highly expressed as tumors become more hypoxic, an event that would be expected as hypoxia is associated with oxidative stress and the NADPH synthesized by the oxidative arm of PPP is upregulated to rescue tumors from death (95).

3.3.2 Overall Differential Gene Expression and GO Term Enrichment

We carried out GO term enrichment analysis for the genes that are differentially expressed between normal and tumor in the TCGA-LIHC dataset. As hundreds of GO terms were enriched especially for biological processes, many genes were found in more than one GO term, thus further filtering was done to remove redundant GO terms as explained before. Figure 3.23 shows the main GO terms enriched in tumor and normal tissues. In almost complete agreement with microarray data, we observed that tumors were mostly enriched in genes involved in DNA replication processes, nuclear morphology and amino acid transport. On the other hand, genes downregulated are involved in essential liver functions such as lipid catabolism, gluconeogenesis and its regulation, and bile acid metabolism. In addition, the retinol pathway is well-known to be downregulated in many tumors. Surprisingly, some of the most enriched terms were related to immune response, and these data are in good agreement with data reported on the suppression of immunity in liver tumors (150). This is an interesting area especially after the promising success of immunotherapeutic approaches in many tumors, including liver (56).

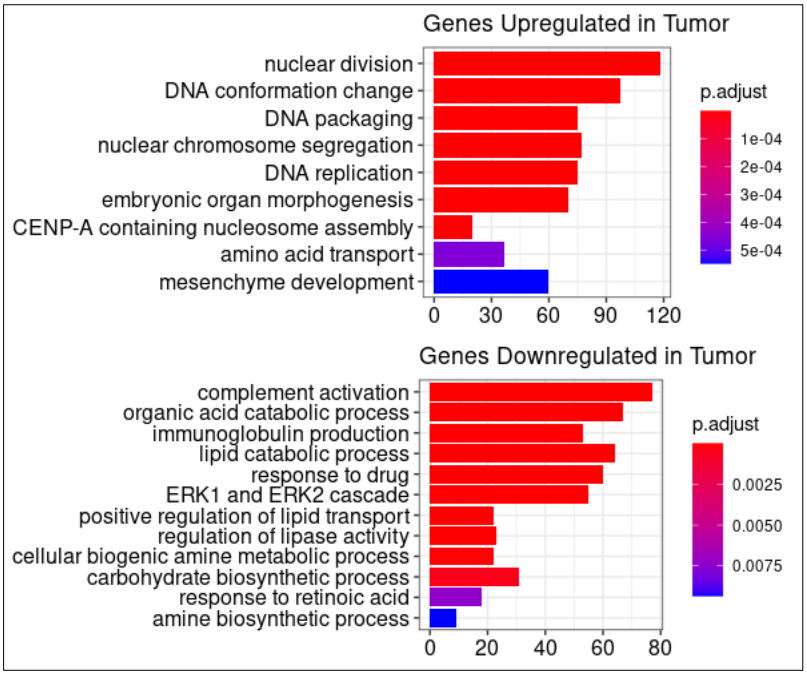

Figure 3.23 BP GO Terms Enriched for Genes Significantly Upregulated and Downregulated in LIHC.

The numbers at the bottom of each graph show the number of genes for each GO Term.

Since many of the enriched processes were enzymatic, the GO term enrichment for molecular functions (MF) was carried out. This is a separate category that classifies genes according to their molecular functions in the cell such as binding certain molecules, transportation, detoxification etc. The results of this analysis is shown in Figure 3.24.

MF GO terms enrichment shows that besides the conventional extracellular matrix and DNA replication-related processes, a significant number of genes upregulated in tumors were members of ion channel family, an important group of gated channels whose function depends on mechanical, chemical and electrical stimuli. These channels are essential for normal cellular physiology and they have been found to regulate cellular apoptosis, migration and proliferation, all of them important hallmarks of cancer (151). MF GO terms enriched in normal tissue on the other hand

were related to immunity, lipid catabolism, bile acid metabolism (tetrapyrrole), monooxygenase enzyme family and amino acid catabolism, especially urea cycle which is one of the vital liver functions.

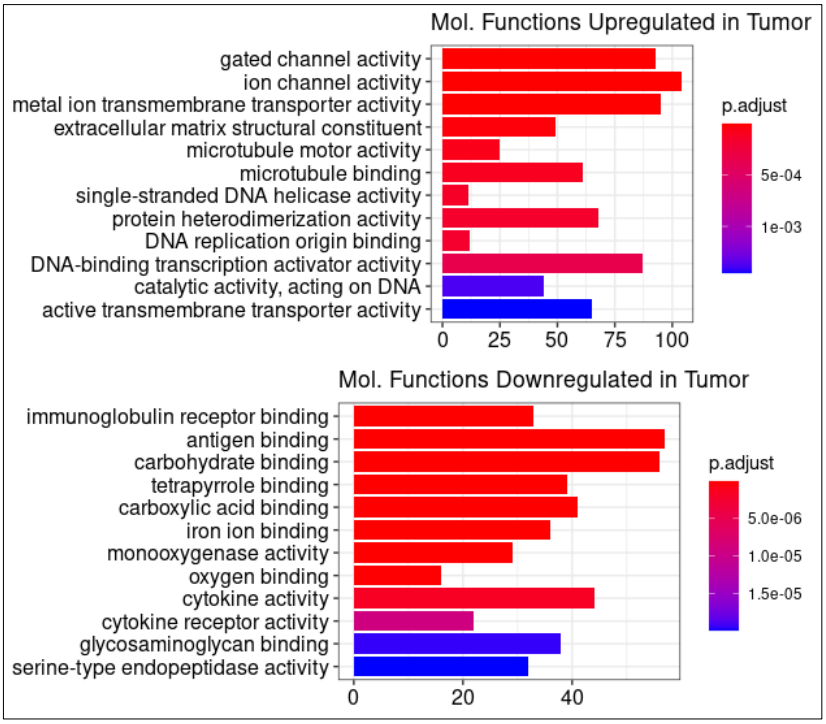

Figure 3.24 MF GO Terms Enriched for Genes Significantly Upregulated and Downregulated in LIHC.

The numbers at the bottom of each graph show the number of genes for each GO Term.

3.3.3 Metabolic Genes Differential Expression

When GO terms output was analyzed more carefully, we observed the enrichment of several metabolic processes. However, most of the pathways tend to lose statistical power because of the large number of genes selected for enrichment, majority of them related to GO terms with hundreds of genes such as cell division, DNA replication etcetera, thus metabolic pathways with few genes did not get enriched. Since the main focus of this study is metabolism in tumor, besides overall DGE, we

evaluated in more detail the expression of the enzymes identified in section 2.10. Following a similar approach, 2424 enzymes were identified in TCGA-LIHC dataset and their expression in tumor versus normal was analyzed. In total, 1497 enzymes (~60% of all) showed significant differential expression between tumor and the adjacent normal tissues (FDR < 0.05), 513 of which were chosen for the more stringent criterion of more than 1LFC. Among them, 264 were upregulated and 249 were downregulated in tumors, a surprisingly balanced distribution. The expression of these enzymes is shown as a heatmap in Figure 3.25.

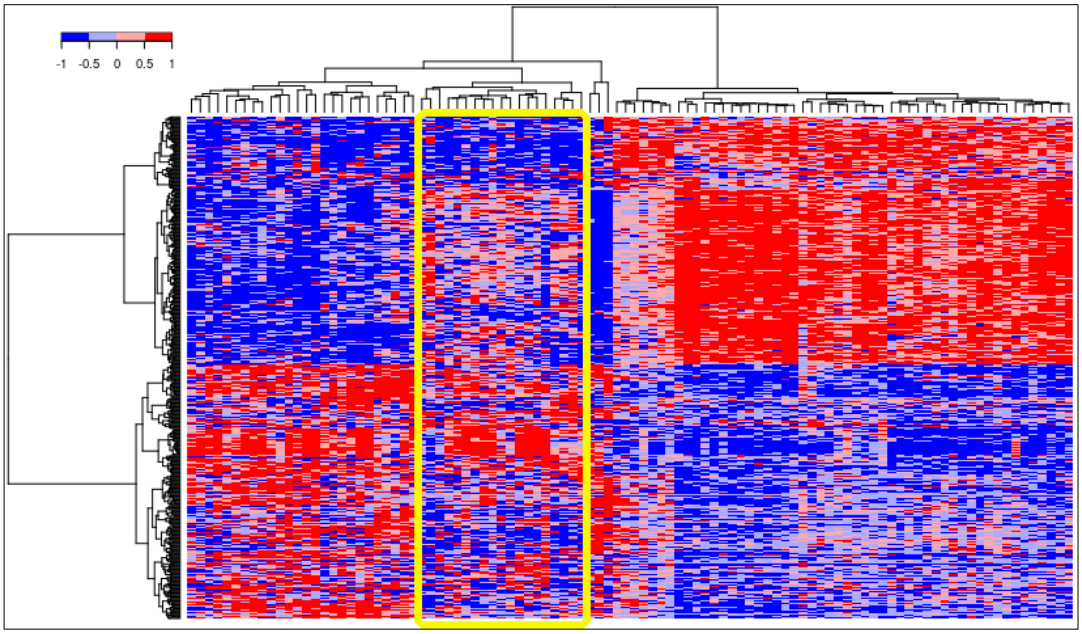

Figure 3.25 Significantly Differentially Expressed Enzymes between Tumor and Normal LIHC Samples.

The cluster on the right shows tumor samples, and the cluster on the left normals. The yellow box in the tumor cluster shows one of the two main subclusters which were further analyzed. Heatmap was constructed with Euclidean distance and Ward.D2 linkage. Color bar at the top shows expression changes in LFC.

Both tumors and normal samples were almost completely separated into two major clusters. In addition, while the expression of genes is almost uniform in normal

samples except a small outgroup, they form two separate subclusters in tumors (highlighted in yellow on the heatmap) which are quite different from each other. This shows that tumor samples have a lot of heterogeneity, not an unexpected feature especially for liver cancer, and that they may have profound transcriptional and mutational differences that may lead to different clinical outcomes.

In the first subcluster, the expression of enzymes is stronger and more uniform, irrespective of whether they are down- or upregulated as compared with the samples highlighted by the yellow box. Of note, AKR1B10 clustered with TKT, G6PD, ME1, TXNRD and UGT enzymes in the first subcluster, and their expression is higher in tumors. When individual genes constituting the GO terms enriched in normal samples were checked individually, we observed that they are involved in many normal physiological liver functions, including fatty acid oxidation in both mitochondria and peroxisomes. We next examined the expression of Cytochrome P450s (CYPs). Out of 28 CYPs that showed significant DGE between normal and tumors, 22 were downregulated and 6 were upregulated in tumors. The 6 upregulated CYPs included CYP7A1, CYP17A1, CYP19A1 (sterol metabolism), CYP27B1, CYP2R1 (vitamins) and CYP2F1 (xenobiotics).

The downregulated CYPs are mostly involved in the metabolism of eicosanoids, fatty acids and vitamins (152). CYP540s are important enzymes involved in drug metabolism, especially xenobiotic detoxification, cholesterol and bile acid metabolism, as well as lipid modifications, changing many signaling pathways or even the lipidome composition of tumors where they are suppressed, overexpressed or mutated (152). They harvest the reducing electrons from NADPH and transfer them to their substrates which get reduced generally to less toxic forms. However, there are also many other cases when CYP450s act as activators of carcinogens, especially certain CYP450 genetic variants (153). The suppression of nearly 50% of all known CYP450s in liver tumor, majority of them involved in lipid metabolism could be important and deserves more attention. Lipids are important components of

plasma and other organelle membrane makeup, and as a result may have huge implications on cellular physiology such as membrane potential which can drastically affect cellular bioenergetics, survival and function, cell motility and apoptosis.

We also observed that the expression of aldehyde/alcohol dehydrogenases (ADH/ALDH), which reduce NAD^+ during their enzymatic reactions to produce NADH were downregulated. The only exception was ALDH3A1 which uses $NADP^+$ as co-factor instead of NAD^+ (154). NAD-reducing retinol dehydrogenase 16 (all-trans) (RDH16) and xanthine dehydrogenase (XDH) had the same fate as well. Another class of enzymes significantly downregulated in tumor samples were flavoproteins, or those using flavin adenine dinucleotide (FAD) as a cofactor for oxidizing their substrates and generate its reduced form (FADH). These enzymes include dimethylglycine dehydrogenase (DMGDH), glutaryl-CoA dehydrogenase (GCDH), isovaleryl-CoA dehydrogenase (IVD), and phosphoglycerate dehydrogenase (PHGDH), in addition to Acyl-CoA dehydrogenase (ACADs). This phenomenon can be speculated to be either due to hypoxia and the inability of cells to be supplied with enough oxygen, therefore inhibiting their ability to metabolize fatty acids. Another reason could be the decreased levels of Succinate Dehydrogenase (Complex II, shown in pan-cancer metabolism section) in mitochondria which oxidizes FADH to FAD via ACAD in the first step of beta oxidation, making it unavailable for another reaction cycle.

Besides the classical kinases involved in cell cycle and DNA replication, enzymes involved in lipid biosynthesis were significantly upregulated in tumors as shown in section 1 of this chapter (Figure 3.2). Additionally, members of AKR family (AKR1B10, AKR1C3) and PPP (G6PD, TKT and TALDO1 (below 1 LFC)) were upregulated. The enzymes involved in glycolysis were also drastically upregulated while the expression of those involved in gluconeogenesis was suppressed (Figure 3.1). This may explain the mechanism by which the expression of ADH/ALDHs was

decreased, while that of AKRs was increased. Since increased aerobic glycolysis (Warburg effect) increases the pool of NADH in the cytosol, it can lead to suppression of reactions carried out by NAD-reducing ADH/ALDH enzymes by mass-action, while enhanced PPP, which reduces $NADP^+$ into NADPH can lead to increased expression of NADPH-oxidizing AKR family members as well as ALDH3A1. The mechanism is shown graphically in Figure 3.26.

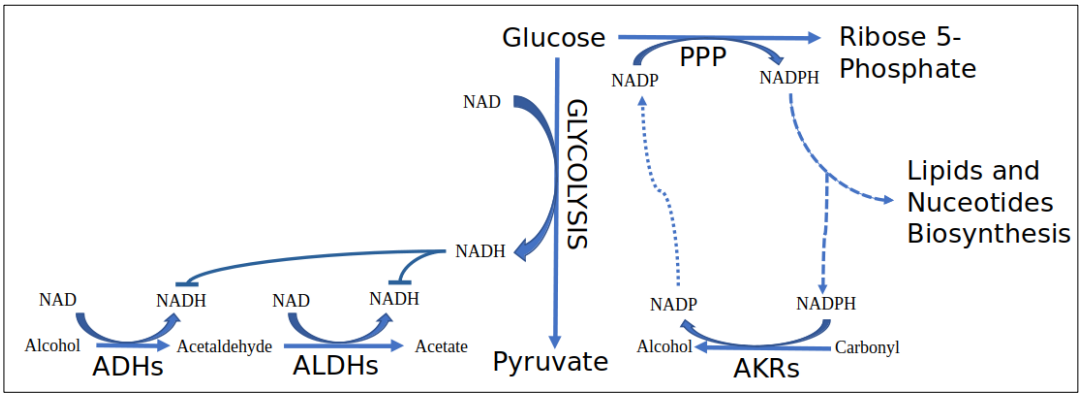

Figure 3.26 Proposed Mechanism for Suppressing ADH/ALDH and Enhancing AKR Expression in LIHC.

Arrows show the direction of reactions while blunt arrows show inhibition. Dashed lines show our proposed interaction between AKRs and PPP.

3.3.4 Differential Gene Expression Analysis of Tumor Subclusters

As shown in the heatmap in Figure 3.25, we obtained two major subclusters in tumor samples and more importantly, the expression of AKR1B10 as well as other PPP genes was higher in the first subcluster. Given that LIHC is well-known to be highly heterogeneous, deeper analysis of such clusters can provide a better picture from both a metabolic and possibly clinical perspective. Therefore, tumor samples from each cluster were retrieved and analyzed separately with their matched normals and evaluated for DGE. As expected, the expression of AKR1B10 as well as many other genes involved in biosynthetic and detoxification reactions were drastically different

between tumor and normal samples. Figure 3.27 shows a summarized picture of the enzymes showing significant differential expression (FDR < 0.05, 1LFC) for both between cluster 1 and 2 (highlighted in yellow in Figure 3.25) with their respective normal tissues, as well as between the clusters themselves.

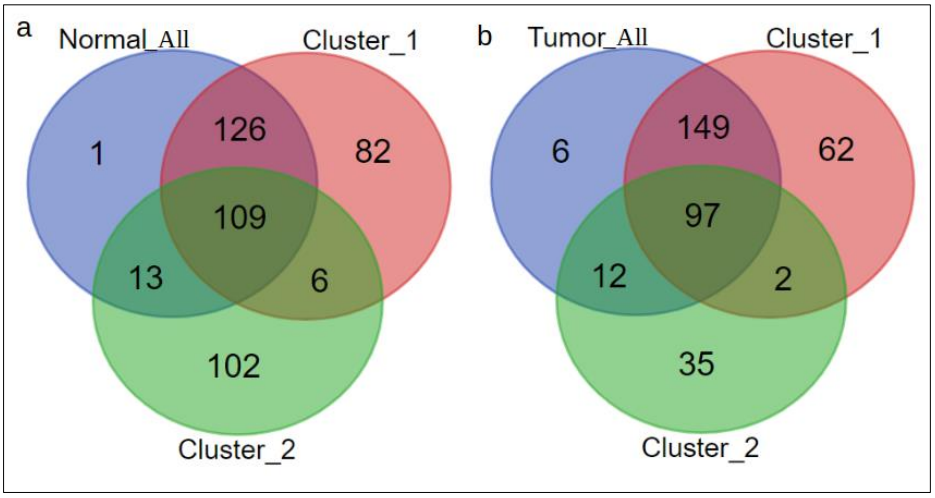

Figure 3.27 Venn Diagrams Summarizing DGE Between Tumor Clusters and Matched Normals.

(A) Genes downregulated in tumor (upregulated in normal). **(B)** Genes upregulated in tumor. Normal_All and Tumor_All stands for 50 samples, while Cluster_1 and Cluster_2 were defined above. Venn diagram was generated in Bioinformatics & Evolutionary Genomics Server (http://bioinformatics.psb.ugent.be/webtools/Venn/).

As was expected from the heatmap, cluster 1 had more uniform gene expression, thus there were more than double up-and downregulated genes with respect to their matched normal tissues as compared to cluster 2. In cluster 1, the profile of up- and downregulated enzymes was similar to that of entire set of 50 samples, however the scale of change (in LFC) was more dramatic, as was expected from the sharp boundaries on the heatmap in Figure 3.25. For example, DGE of AKR1B10 changed from 3.8 to 4 LFC and that of G6PD was even more dramatic from 2.29 to 2.83 LFC (Figure 3.28). Additionally, more AKR family members such as AKR1B1 and AKR1C2-3 showed more than 1 LFC in this cluster. On the other hand, enzymes

involved in betta oxidation, ADH/ALDHs and CYPs showed a more dramatic decrease in cluster 1 when compared to all 50 tumors taken together.

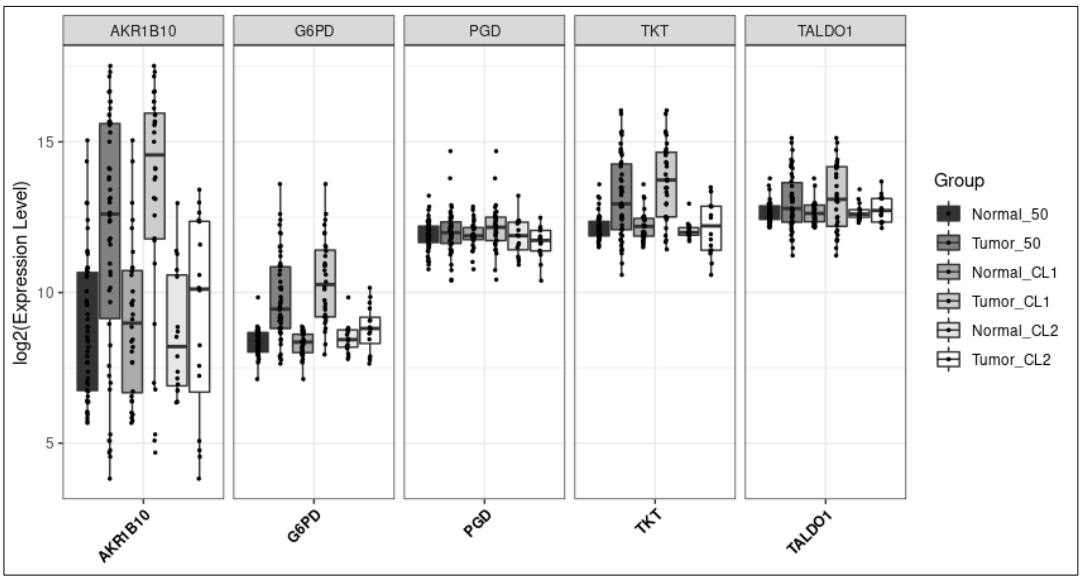

Figure 3.28 AKR1B10 and PPP gene Expression by Clusters.

Normal_50 and Tumor_50 show the expression of the genes for all 50 samples of tumor and normal tissues. Normal_CL1 and Tumor_CL1 show expression of genes in cluster 1 tumor and matched normals, and Normal_CL2 and Tumor_CL2 show the expression of genes in cluster 2 tumor and matched normals. Significance bars are not shown for clarity, however none of the genes showed significant change between cluster 2 tumors and their matched normals. In addition, PGD had no significant change between normal and tumor for all samples, but it showed significant change between Tumor_CL1 and Tumor_CL2.

When DGE between cluster 2 and its normal sample was compared, none of AKR family members (except AKR1C3) or PPP genes were significantly upregulated. In addition, some of the enzymes involved in fatty acid oxidation did not show any significance, and for those that did, all had lower LFC values as compared to cluster 1 tumors matched to their normal samples. Cluster 1 also had stronger glycolysis as shown by the expression of some glycolytic enzymes such as Hexokinases (HK1/2), Glyceraldehyde 3-Phosphate Dehydrogenase (GAPDH), Enolase (ENO2) and Pyruvate Kinase (PKM/P). On the other hand, some CYP450s and enzymes involved

in beta oxidation were significantly higher in cluster 2 as compared to 1. We can therefore conclude that cluster 1 with higher PPP and more glycolysis has larger pools of NADH and NADPH, and therefore less expression of NAD^+ reducing (ADH/ALDH) and more expression of NADPH-oxidizing (AKR) enzymes, further supporting the model proposed in Figure 3.25.

Figure 3.27A shows that 109 enzymes were downregulated in all three classes of normal samples (Normal, Cluster 1 and Cluster 2). These genes mostly act as oxidoreductases, such as aldehyde/alcohol dehydrogenases, CYP450s (Class 1-4), Acyl-CoA and Retinol Dehydrogenases among others, fatty acid metabolism such as enzymes of beta oxidation and ketone body synthesis. Additionally, genes for amino acid metabolism, especially enzymes that catalyze the breakdown of branched-chain amino acids were also seen in this list. The second Venn diagram (Figure 3.27B) shows 97 common enzymes upregulated in all tumor groups. These genes are mostly involved in DNA replication and repair, cell cycle checkpoint control and glycosylation. As in Figure 3.27A, cluster 1 showed more commonalities with all 50 samples and less with cluster 2. Of note, most of the AKR family enzymes together with PPP enzymes were found in cluster 1.

To conclude, DGE analysis showed that the expression of genes in liver tumors and matched normal adjacent samples could be separated into two clusters which have commonalities and differences with each other. The genes of interest encoding AKR1B10 and PPP enzymes were expressed at significantly higher levels in cluster 1 as compared to the overall tumor-normal and cluster 2-normal cases (Figure 3.28). A link between AKR1B10 and PPP has not been reported in the literature yet; however, it in light of the results above as well as the biochemistry of AKR1B10 which requires reducing electrons from NADPH for its enzymatic functions, and PPP which generates NADPH, the idea that they cross-talk is quite plausible and it is highly likely that a major branch of glycolysis such as PPP drives the expression of AKR1B10 and/or other AKR family enzymes in LIHC.

3.3.5 Genomic Analysis for 50 Samples

Since mutations and genomic instability are one of the hallmarks of cancer (7), tumor samples were also evaluated for single nucleotide changes, insertion/deletion (indel) mutations, as well as major genomic rearrangements, also called copy number variations (CNV) such as amplifications and deletions. The analysis for this section were carried out on the cBioportal webserver through a number of steps (155,156). Patient IDs representing the tumors matched to normal samples were uploaded on the server, analyzed for mutation enrichment, the results were downloaded and the most frequently mutated genes were selected by mutation frequency and re-uploaded to the webserver to create oncoprints which show mutation type and distribution among samples. The oncoprint in Figure 3.29 shows the top 15 genes that had mutation frequency above 10% among the 50 tumor samples.

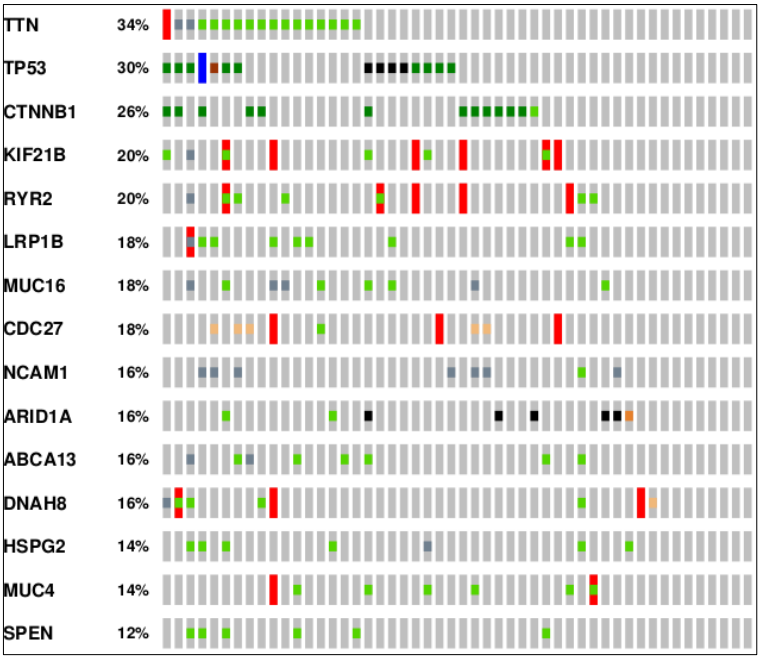

Figure 3.29 Oncoprint of the Most Frequently Mutated Genes in 50 Tumor Samples. The figure shows various types of mutations, the most common being single nucleotide changes missense mutations (dark and light green), gene amplification (red) and truncating mutations (black).

Mutation analysis show that Titin (TTN), mainly a muscle protein which has recently been found to be involved in chromosomal structure (157), TP53, one of the most frequently mutated genes across all tumors (158), and Beta Catenin (CTNNB1), an important co-activator downstream of the oncogenic wingless signaling pathway (159) were the most frequently mutated genes. The high number of mutations in TTN may be completely due to chance since the protein is composed of over 30000 amino acids, increasing its chance of mutations substantially (157). Most of the mutations are of single nucleotide missense type, but no significant changes in survival or gene expression was observed when samples were compared based on mutations of these genes except for higher mutation burden on samples with mutated TP53 ($n = 17$) compared to the wild-type ($n = 33$) (Kruskal-Wallis, $p = 0.016$). This is expected as the p53 protein functions as a tumor suppressor.

Another important set of genes whose mutation profiles were specifically evaluated included the CYP450s, ADHs, ALDHs, AKRs and PPP enzymes because we observed that their expression was significantly changed between normal and tumor samples. It is well-known that such enzymes, especially CYP450s (160) and ADH/ALDHs (161) are prone to mutations, especially SNPs which drastically change their enzymatic rates and alter their substrate specificity. These changes have also been associated with various diseases such as cancer. We examined these genes for mutations in cBioportal by uploading the enzymes on the server, but no mutation of any nature above 4% (2 samples) frequency was detected (data not shown). We have therefore discarded the mutation factor from our analysis.

We next evaluated the enrichment of CNVs in the 50 tumor samples. Following the same procedure as above, tumor samples were screened for genomic alterations in cBioportal server, and the most frequently altered genes were re-uploaded in the server to create oncoprints. The data show that amplification was the most common form of genomic alteration (Figure 3.30).

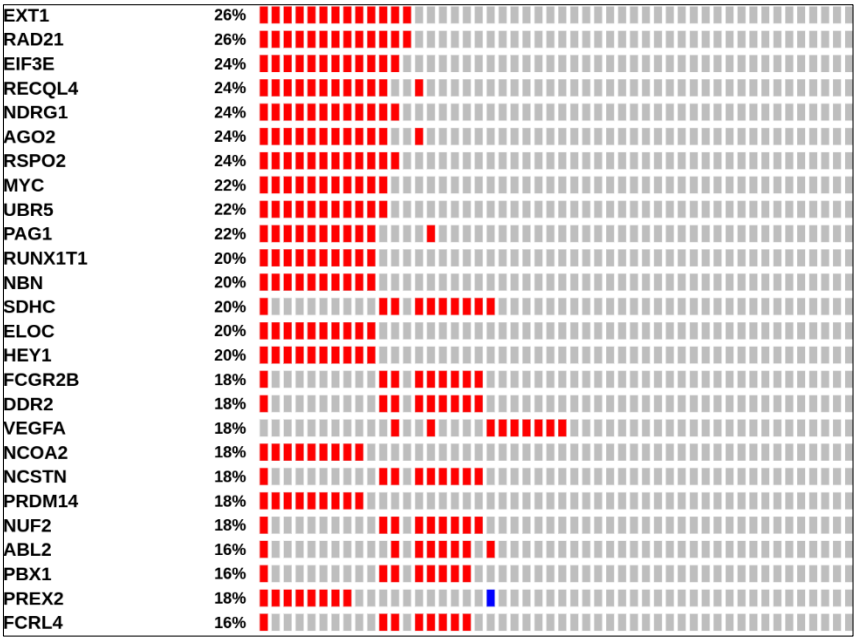

Figure 3.30 Genomic Alterations Oncoprint.

The top 24 genes showing more than 15% genomic alterations are shown. Red boxes represent amplifications, blue deletions, and grey no alteration.

Most of the genes shown in Figure 3.30 were amplified and very few of them had deletions. In addition, the majority of these genes are involved in DNA replication and repair. There was no significant upregulation of these genes between CNV positive and negative samples except RAD21 (FDR < 0.05, LFC < 1) which is a well-known DNA repair enzyme and it is associated with earlier tumor recurrence in breast cancer. Therefore, amplification of this gene makes it an obvious candidate in tumorigenesis (162). On the other hand, EXT1 is a glycosyl transferase enzyme involved in heparan sulfate biosynthesis, however studies have shown that the protein can act as a tumor suppressor in leukaemia (163). EIF3E is a well-known translation initiation factor and studies have associated it with poor prognosis in some tumors, particularly esophageal squamous cell carcinoma (164). The expression of the amplified genes in CNV positive patients was also compared to their matched normal counterparts, and surprisingly some of them had significantly

lower expression in tumor (FCGR2B, IL10, PBX1, PTPRC, MYC and NTRK1), while majority showed no significant difference at all.

Next, the expression of AKR1B10 and PPP genes in the cluster of 20 patients showing enrichment of multiple gene amplifications (CNV positive) was compared to other 30 tumor samples with no or very sporadic amplifications (CNV negative). All of the genes evaluated, except AKR1B10, had significantly higher expression in the CNV positive samples (Figure 3.31).

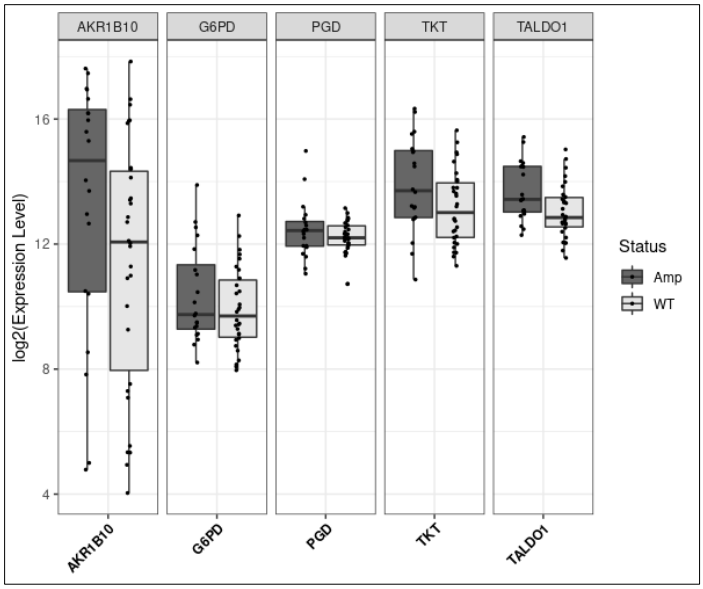

Figure 3.31 AKR1B10 and PPP Gene Expression in Samples with Amplification or No Genomic Changes.

All the genes except AKR1B10 showed significant differences between the two conditions (Wald's Test, $p < 0.05$). Amp represents CNV positive and WT represents CNV negative samples.

Given the function of PPP as a source of nucleotides as well as NADPH and the fact that more cell replication leads to further genome instability in tumors, the idea of an association between PPP and genome instability is very plausible. However, this needs to be verified in controlled settings or large genomic studies which are beyond

the scope of this thesis. Interestingly, when overall survival differences between CNV positive and CNV negative patients was compared, we saw a significant difference of 30 months (median survival 21 and 51 months, respectively) with longer survival in CNV negative patients as shown by the Kaplan-Meier (KM) plot (Figure 3.32).

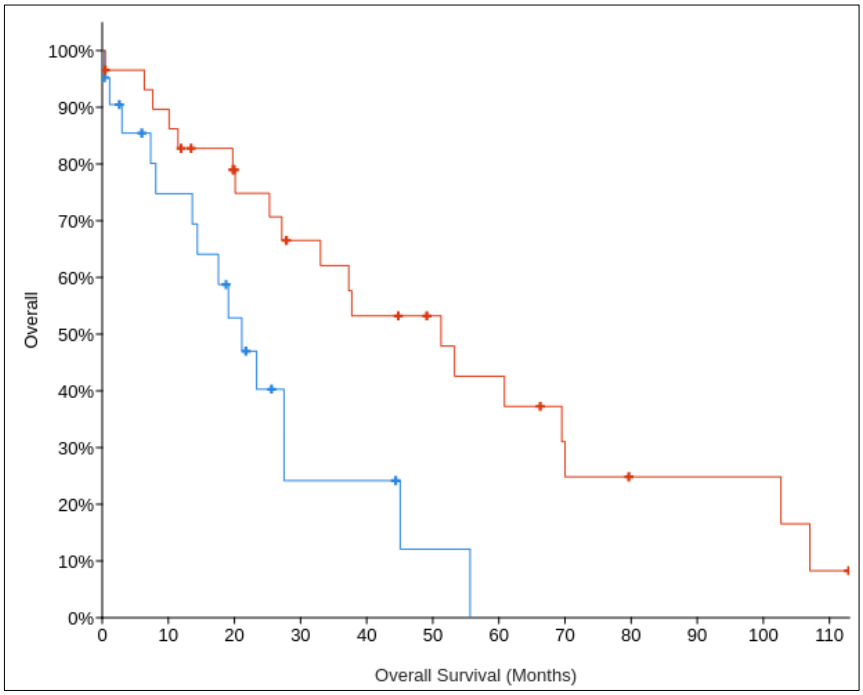

Figure 3.32 Kaplan-Meier Plot of CNV Positive and CNV Negative Patients.
The blue line shows CNV negative and red line CNV positive patients. Y-axis shows the percentage of patients' survival over time and x-axis overall survival in months (Log-rank p-value < 0.01). The plot was generated on cBioportal server (155,156).

A median survival difference of 30 months is highly significant for LIHC because most of the current therapies do not extend patient OS by more than a year (Section 1.5.3). In order to better understand the general molecular mechanisms, DGE between CNV positive and negative samples was carried out. Overall, there were around 500 genes showing significant change (FDR < 0.05, LFC > 1), 173 upregulated in CNV positive and 337 in CNV negative (Figure 3.33).

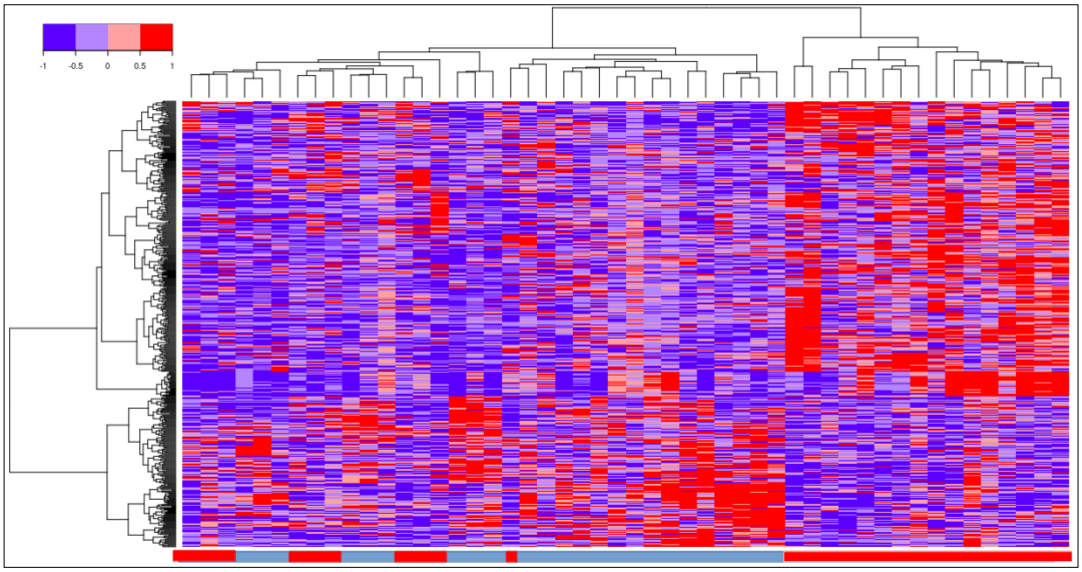

Figure 3.33 Heatmap of Significantly Differentially Expressed Genes Between CNV Positive and Negative.

Only genes with 1LFC and FDR < 0.05 were selected. Heatmap constructed with Euclidean Distance and Ward.D2 linkage. Red columns represent CNV negative and blue cluster CNV positive samples.

We next classified the genes according to biological pathways or GO terms to understand any specific process that could be attributed to such a remarkable difference in patient survival. However, the CNV positive genes did not form any specific cluster. On the other hand, genes upregulated in CNV negative were mostly involved in immunological response (Figure 3.34). Based on this information, it can be hypothesized that CNV positive patients may have shorter survival because they have higher glycolysis and PPP, suggesting more biosynthesis, more cellular replication and thereby genomic instability. Additionally, we can also speculate that CNV negative patients have stronger immunological response against the tumor, thus extending patient survival.

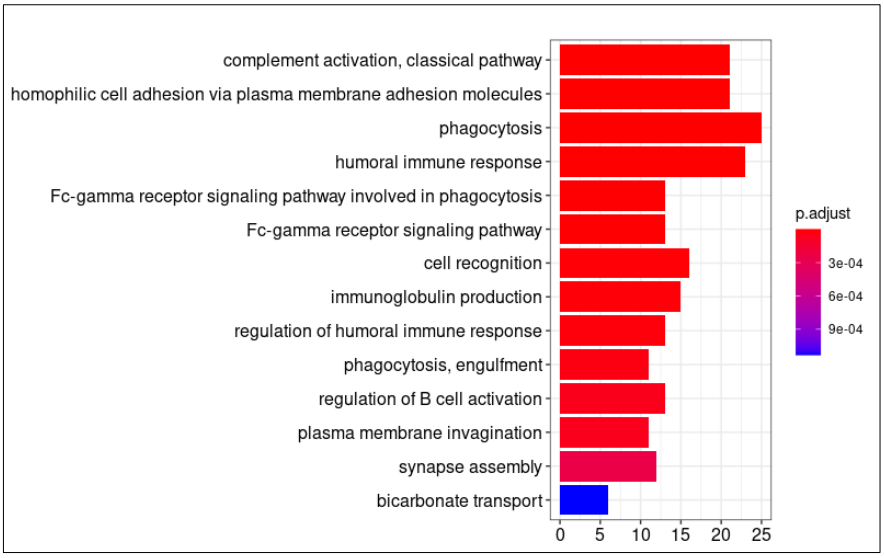

Figure 3.34 GO Term Enrichment of Genes Upregulated in CNV Negative Patients. The GO terms were selected to remove the redundant ones as mentioned previously.

3.3.6 Differential Methylation Analysis

It is now a well-established fact that methylation is one of the most important mechanisms for regulating gene expression, and it is quite widespread in cancers (165). Traditionally, from the perspective of oncology research, hypermethylation leads to downregulation of tumor suppressor genes, and hypomethylation leads to overexpression of oncogenes, thus helping or causing tumorigenicity. In addition, hypomethylation on a global scale leads to genetic instability, which is one of the hallmarks of cancer (6,7). We therefore examined the methylation status of AKR1B10 and PPP genes. For that purpose, we carried out differential methylation expression analysis (DMA) between tumor and their matched normal samples using the data available in the TCGA repository. DMA showed remarkable alterations in global methylation between tumor and normal samples with more than 50% of the probes having significant changes (FDR < 0.05). These changes were identified in both coding and non-coding regions, such as 5'- and 3'-untranslated sites (UTRs), gene body, exons and transcription start sites (TSSs), some with significant changes.

However, since the real function and impact of each of these sites on gene expression is not well-established and subject to debate (166), we concentrated only on the TSS1500. The expression of beta values for AKR1B10 and PPP genes is shown graphically in Figure 3.35.

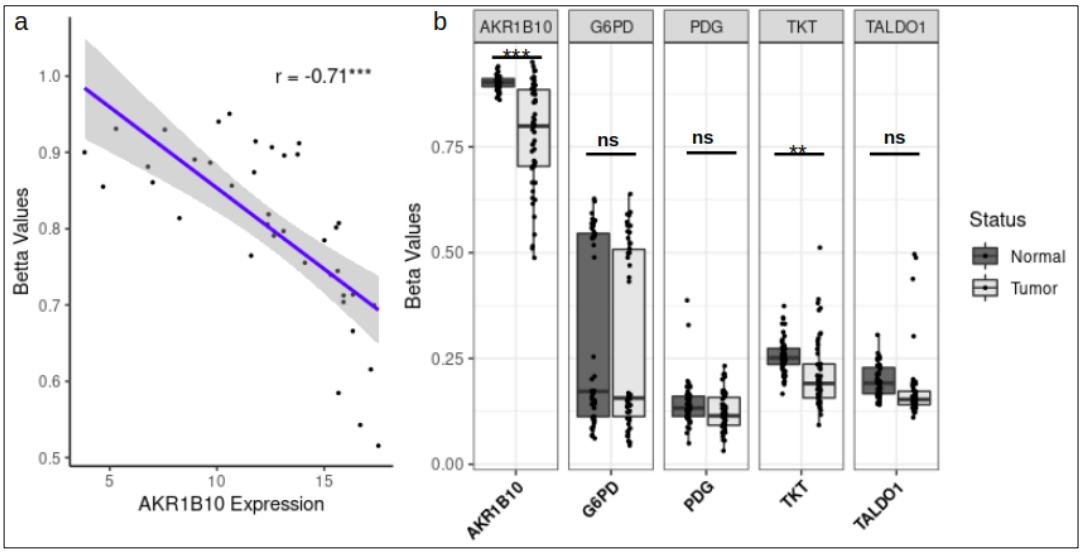

Figure 3.35 Methylation of AKR1B10 and PPP Genes in Normal and Tumor Tissues. **(A)** Pearson correlation of AKR1B10 expression and beta-values for the chosen probes tumor samples. Correlation coefficient is shown inside the plot. **(B)** Probes used in this graph are: AKR1B10: cg11693019; G6PD: cg26799772; PGD: cg12796186; TKT: cg00223877 and TALDO1 cg08037817. Significance levels:* p< 0.05, ** p< 0.01, *** p< 0.001, ns not significant by t-test.

The figure shows a highly significant hypomethylation of AKR1B10 on TSS1500 site in tumors, a state that is negatively correlated with the gene expression. On the other hand, the expression of AKR1B10 had no correlation with methylation in normal tissue. G6PD, PGD and TALDO1 showed no change in methylation status while TKT was hypomethylated in tumor. In addition, there was no correlation between the expression of any of the PPP genes with methylation of TSS1500 site both in normal and tumor samples (data not shown).

To summarize, using tumor and normal matched samples from TCGA-LIHC dataset, we have conformed and refined the results obtained from GSE6764 microarray. We have shown that AKR1B10 and PPP, especially the rate-limiting enzyme G6PD were generally upregulated in LIHC. More particularly, AKR1B10 and G6PD were overexpressed in approximately 80% and 85% of tumors, respectively. We also showed that the expression of these genes was not dependent on tumor grade (except G6PD in stage III).

Overall, if the data from pan-cancer and LIHC are combined, we observed that liver tumors have elevated levels of glycolysis and PPP, which can fuel biosynthetic reactions such as nucleotide and fatty acid biosynthesis, as well as coping with oxidative stress and most likely drug detoxification. On the other hand, there was an overall decrease in catabolism such as decreased fatty acid (beta oxidation) and amino acid oxidation, in particular one of the most essential functions of liver, the urea cycle. Tumor samples were transcriptionally separated in two clusters, the first one showing more drastic changes in the above-mentioned processes than the second.

Single nucleotide mutation analysis showed the enrichment of a number of critical genes being mutated in approximately 30% of the samples. A similar pattern was observed in genomic rearrangements, and we showed that the expression of PPP but not AKR1B10 was increased in the samples with more genomic instability, a state associated with significantly poor overall survival in patients.

Finally, methylation analysis showed dramatic changes in the overall methylome of LIHC compared to normal. We also showed that the expression of AKR1B10 is highly likely to be regulated by methylation, while that of PPP genes was less likely based on methylation status of TSS1500 site.

3.4 AKRB10 High and Low-Expressing LIHC

The status of AKR1B10 as a bona fide liver tumor biomarker has been under dispute for various reasons which were summarized in section 1.6. In this section, we have examined the association of AKR1B10 expression with LIHC and aimed to unravel the underlying molecular mechanisms. For this purpose, 364 patients with overall survival data were chosen from the TCGA-LIHC cohort. Their general characteristics are shown in Table 3.3.

Clinical characteristics of age, gender, grade, TNM stage, Neoplastic Grade and patient overall survival were downloaded from TCGA, while mutation status for 26 most frequently mutated genes were retrieved from published literature (39) and combined together. Afterwards, the association of each characteristic with patient survival was calculated using univariate cox-proportional hazard (log-rank $p < 0.05$). Only the statistics for the three most significantly mutated genes are shown in the table since there was no significance for the others. The presence or absence of HBV and HCV is not shown because their status in the clinical data was not clear despite efforts to use them as clinical variable in patient survival.

3.4.1 Association of AKR1B10 with Patient Survival

After normalization of the RNA-seq read count, LIHC patients with follow-up data were stratified according to AKR1B10 expression first into two groups according to median, and then into 4 groups from which the highest and lowest 25% (first and fourth quartile) were used for analysis. The Kaplan-Meier (KM) Plots for both stratifications are shown in Figure 3.36.

Table 3.3 TCGA-LIHC Patient Characteristics.

No column represents the number of patients for each variable. HR represents hazard ratio with 1 being the reference. 0.25% and 97.5% represent the lower and upper confidence interval for HR. Significance threshold log-rank p-value < 0.05.

Characteristic	*Variable*	*No*	*HR*	*0.25% CI*	*97.5% CI*	*p-value*
Gender	Male	237	1			
	Female	113	0.82	0.57	1.17	0.3
Age	<= 60	167	1			
	>60	183	0.80	0.56	1.15	0.2
Stage	I	166	1			
	II	82	1.49	0.92	2.42	0.11
	III	77	2.55	1.65	3.93	<0.001
	IV	5	4.94	1.76	13.86	0.002
Neoplasm Grade	G1	50	1			
	G2	168	1.26	0.72	2.19	0.417
	G3	115	1.28	0.72	2.29	0.402
	G4	12	1.75	0.64	4.79	0.279
TNM Stage	T1	175	1			
	T2	86	1.49	0.94	2.36	0.09
	T3	40	2.30	1.41	3.76	<0.001
	T3A	26	2.78	1.38	5.57	0.004
	T4	12	5.34	2.67	10.69	<0.001
TP53	WT	236	1			
	MUT	114	1.53	1.056	2.22	0.0246
CTNNB1	WT	253	1			
	MUT	97	1.06	0.7084	1.59	0.776
ALB	WT	302	1			
	MUT	48	1.27	0.458	1.35	0.386

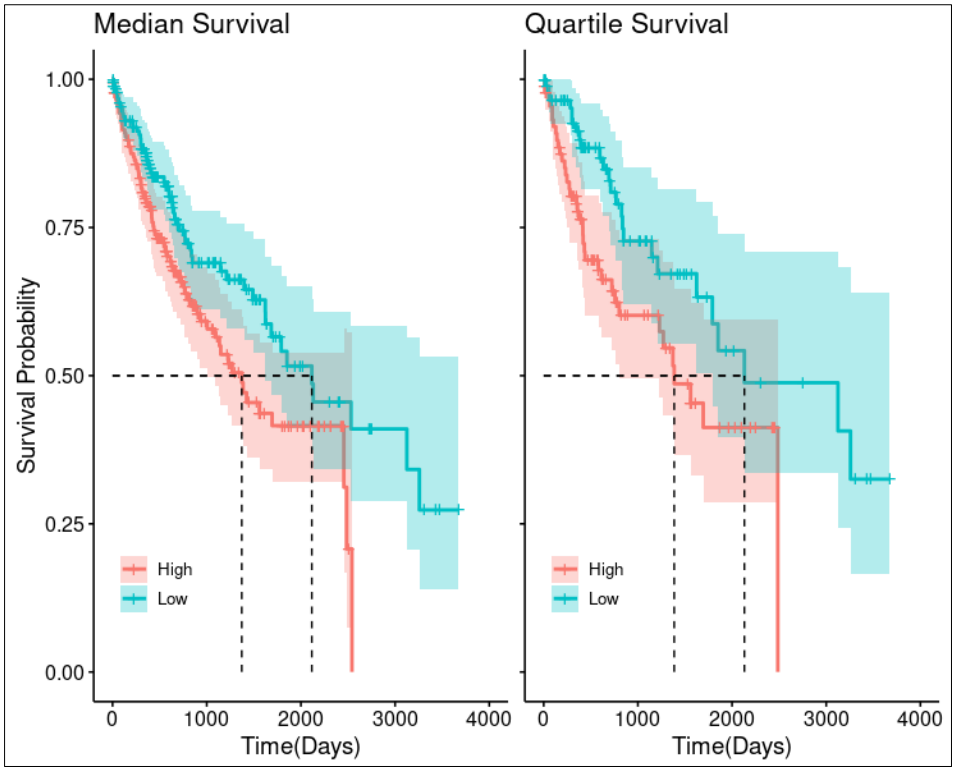

Figure 3.36 KM Plots of Patient Survival According to Median and First and Fourth Quartile of AKR1B10 Expression.

Patients' groups are color-coded and the horizontal dashed line shows median survival. Both show significant survival differences (log-rank p = 0.015 and 0.018, respectively).

Univariate Cox proportional hazard model showed that in both cases patients with higher expression of AKR1B10 ($AKR1B10^{High}$) had significantly lower survival compared to patients with lower expression of AKR1B10 ($AKR1B10^{Low}$). In the case where patients were stratified by the median expression of AKR1B10, hazard ratio (HR) for $AKR1B10^{High}$ was 1.55 (lower confidence interval (LCI) 1.1, upper confidence interval (UCI) 2.2, p-value 0.015). This can be interpreted as a patient with higher than median expression of AKR1B10 is approximately 1.55 times more likely to die than a patient with lower AKR1B10 expression. On the other hand, when the model was fit only on the first and fourth quartile, HR was 1.9 (LCI = 1.11, UCI = 3.18, p-value = 0.018). Thus, a patient whose expression of AKR1B10 is in the fourth quartile is almost twice as likely to die than a patient in the lowest quartile.

Median survival of patients for both models was also similar: in both cases, the median survival of AKR1B10^{Low} patients was more than 2 years longer than AKR1B10^{High} category. The idea for this entire study was conceived from the survival differences shown in this section.

Next, we evaluated the independence of AKR1B10 expression from other clinical and mutation factors (Table 3.4). This time the expression of AKR1B10 was converted into a binary value and the dependence on each clinical factor was determined using Chi-Square test. As the table shows, with few exceptions the dependencies for median or quartile stratification were similar.

After patient stratification according to AKR1B10 expression, exploratory data analysis indicated that using all 364 tumor patients and determining their gene expression differences was not very feasible since liver cancer is very heterogenic. For more relevant results and more transcriptionally homogenous sample grouping, we selected to further analyze the differences in gene expression and methylation between the first and fourth quartiles of AKR1B10 expression, and from now on, the terms AKR1B10^{High} and AKR1B10^{Low} are exclusively referring to them. The following sections will show the results of different bioinformatics approaches which show the potential link between AKR1B10 and PPP in the AKR1B10^{High} group and its implications for liver patient stratification as well as new approaches in managing or treating liver tumors.

Table 3.4 Clinical Dependencies of Patients Stratified by AKR1B10 Expression.

Variables		*Median*			*Quartile*		
Characteristic	Variables	High	Low	p-value	High	Low	p-value
Gender	Male	102	135	< 0.001	72	46	<0.001
	Female	41	72		15	41	
Age	<= 60	78	89	0.24	36	39	0.76
	>60	98	85		51	48	
Grade	I	84	82	0.14	45	40	0.15
	II	45	37		22	19	
	III	31	46		13	23	
	IV	4	1		2	0	
Neoplasm Grade	G1	18	32	0.14	6	19	0.023
	G2	91	77		46	38	
	G3	59	56		33	25	
	G4	5	7		1	3	
TNM Stage	T1	90	85	0.048	46	41	0.022
	T2	49	37		24	18	
	T3	17	23		7	13	
	T3A	7	19		1	8	
	T4	8	4		6	2	
TP53	WT	105	131	0.003	49	66	0.01
	MUT	71	43		38	21	
CTNNB1	WT	124	129	0.52	59	66	0.31
	MUT	52	45		28	21	
ALB	WT	147	155	0.18	70	80	0.048
	MUT	29	19		17	7	
AZIN1	WT	132	159	< 0.001	63	82	<0.001
	MUT	44	15		24	5	
KEAP1	WT	160	172	0.002	77	86	0.013
	MUT	16	2		10	1	
NFE2L2	WT	156	170	0.002	76	85	0.021
	MUT	20	4		11	2	

3.4.2 Gene Set Enrichment Analysis Between AKR1B10^{Low} and AKR1B10^{High} Expressing Samples

First, we used gene set enrichment analysis (GSEA) to evaluate enriched biological processes and pathways in AKR1B10^{High} and AKR1B10^{Low} to better understand the differences seen in patient survival. For this purpose, RNA-Seq read counts normalized and VST-transformed by *DESeq2* package. Expression and phenotype files were prepared according to instructions of the software and the analysis was run on GSEA's desktop application with default parameters. The reason for choosing GSEA is its robustness in capturing whole pathways because of its unique approach which differs a lot from DGE followed by GO term enrichment. It searches for pre-defined and curated gene sets, and based on their correlation, it uses permutations to further refine the results. This enables the identification of connections and enrichment of pathways including genes with modest changes between conditions. This would not be possible by conventional GO term enrichment approaches as they are based often times on arbitrary LFC and FDR cut-off values (167). In order to obtain more comprehensive results, 4 different enrichment methods in GSEA were used: gene ontology, hallmarks, KEGG pathways and Reactome. The results for each of these approaches are shown in detail in the following subsections.

3.4.2.1 GSEA with Gene Ontology

To classify enriched genes according to gene ontology, GSEA was run on C5 molecular signature database (MSigDB) with a minimum set of 15 genes per GO term. Overall, 52 GO terms were enriched (FDR < 0.25) (Appendix A). Figure 3.37 shows the enrichment plots for some of the GO terms. Most terms shown are involved in NADPH production by Pentose Phosphate Pathway, Isocitrate Dehydrogenase 1 (IDH1) and malate dehydrogenase, which is more commonly known as Malic Enzyme 1 (ME1), as well as NADPH consuming pathways such as glutathione and AKR enzymes.

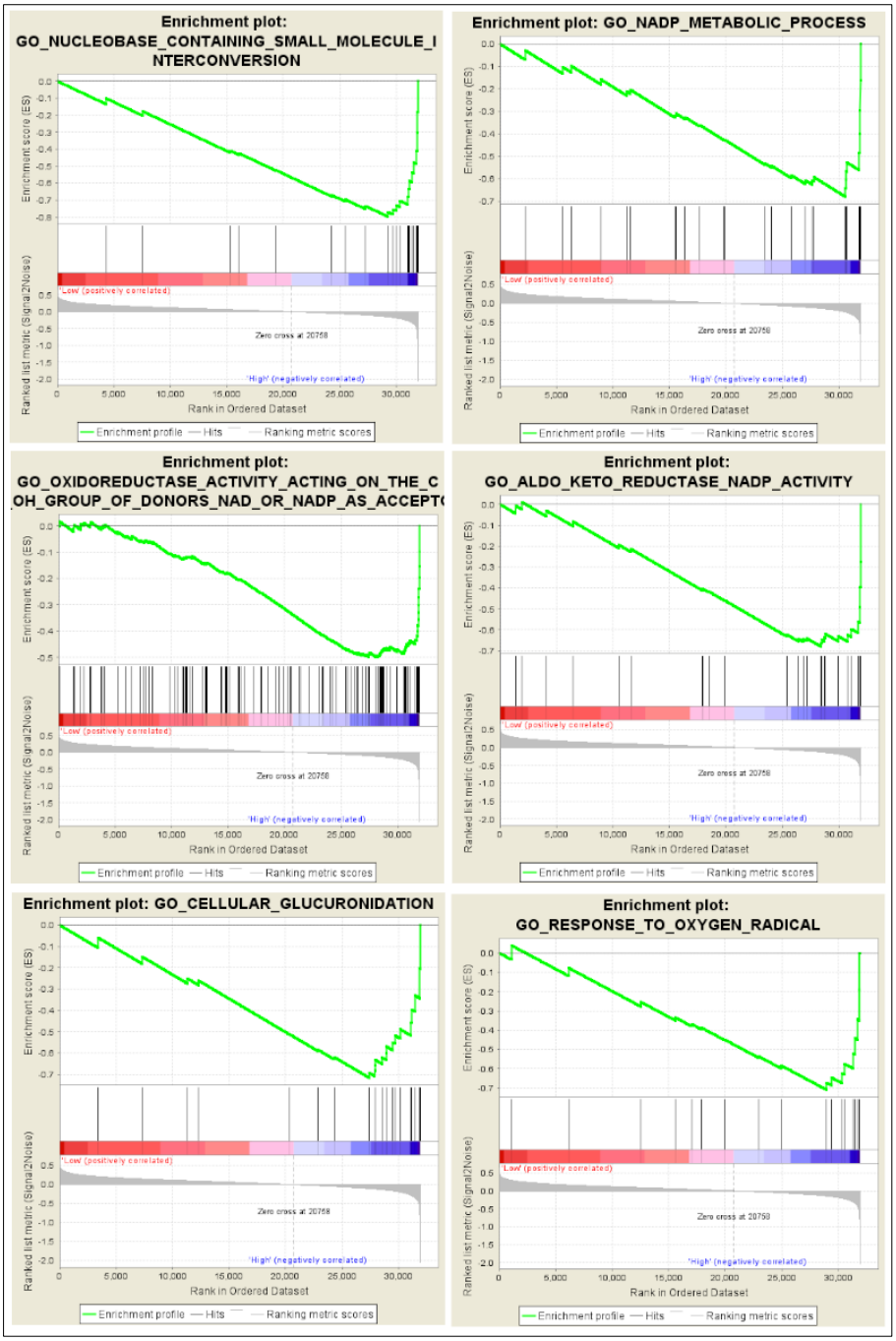

Figure 3.37 GSEA Enrichment Plots Representing NADPH and AKR-Related GO Terms.

These enrichment plots were selected from the statistically significant GO Terms generated by GSEA. The leading edge points downward because of the manner the conditions were fed into GSEA.

IDH1 and ME1 can become important contributors in the NADPH pool of the cell (33) while IDH1 enzyme when mutated, reduces Isocitrate to 2-Hydroxyglutarate (D-2HG) instead of αKG. D-2HG is an oncometabolite commonly detected in gliomas and myeloid leukaemia (96). D-2HG is known to inhibit the enzymatic actions of many αKG-dependent dioxygenases, such as histone demethylases, thus causing major epigenetic changes in cells harboring such mutations. We carried out mutational analysis of IDH1 in our samples and found it to be mutated only in 4 samples out of 180 (3 in AKR1B10 high and 1 in AKR1B10-expressing patients). At the same time, in the original article analyzing the TCGA-LIHC patients, no significant enrichment of IDH1 mutation was found despite the efforts of the research team to create an IDH1 signature (39).

Other important genes enriched by GSEA include multiple members of AKR family, in particular AKR1B1, AKR1C1-4, ALDH3A1 and DHRS3. Brief information about the function of these proteins is provided in chapter 2, while DHRS3 reduces retinols and steroids in the presence of NADPH. Another important group of enzymes enriched in AKR1B10^{High} are those involved in oxidative stress response such as thioredoxin reductases (TXNRD), superoxide dismutates (SOD), quinone (NQO1), glutathione reductases (GSR), glutathione peroxidase (GPX) etc. While high levels of Reactive oxygen species (ROS) are considered to be detrimental for cell survival, low doses of ROS are necessary for normal physiological functions of the cell (168). ROS levels are generally high in cancer because of hypoxia and high metabolic rate to sustain cell division, causing massive damage to DNA, lipids and other biomolecules, and if they are not neutralized, they can cause cell death (169).

One of the main pathways in neutralizing ROS is the glutathione system which utilizes NADPH to reduce ROS, while peroxidized lipids are neutralized by AKRs. The expression of all of these genes in AKR1B10^{High} patients is a strong indicator that besides the role of NADPH in biosynthesis, co-expression of AKR, PPP and Glutathione system proteins makes it possible for the tumors to control oxidative stress and metabolize chemotherapeutic drugs more efficiently. Other important GO terms enriched in AKR1B10^{High} samples were related to amino acid metabolism, in

particular the polyamine pathway and an overexpression of multiple proteasome subunits. This may suggest an enrichment of targeted protein degradation for amino acid utilization in AKR1B10^{High} tumors.

Another important group of enzymes enriched in AKR1B10^{High} tumors are the UDP-glucuronosyltransferase enzymes (UGTs), a family of 22 genes that are involved in the detoxification of drugs and other toxic molecules by attaching a sugar moiety from UDP-sugar donors (170). This conjugation, besides deactivating the molecules, it also makes them more soluble and easier to be excreted through urine. Higher expression of these enzymes has been associated with drug resistance and shorter OS/DFS in multiple tumors, mostly because of their involvement in chemotherapeutic drug detoxification and elimination (171). However, recent studies have also shown that UGTs may be involved in other, yet unknown functions such as oncogenic signaling through their role in metabolizing endogenous bioactive molecules such as bioactive lipid species and steroid hormones (172).

Only 2 GO terms were enriched in AKR1B10^{Low} patients, both of them representing genes involved in DNA methylation such as DNA methyltransferases (DNMTs) and MECP2 (Figure 3.38).

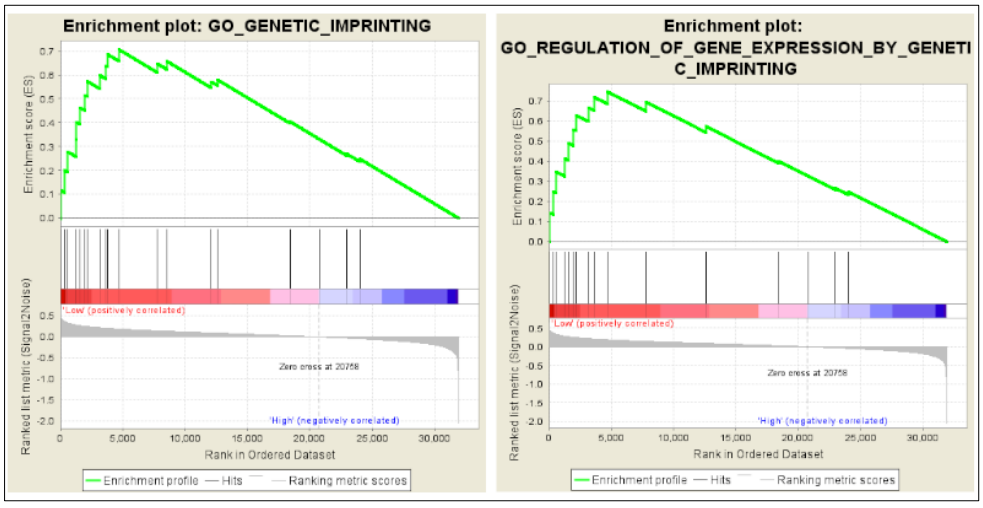

Figure 3.38 GSEA GO Term Enrichment Plots for AKR1B10^{Low} Patients.

Interestingly, these GO terms also included insulin-like growth factor 2 (IGF2) and glycogen synthase kinase 3 beta (GSK3B). When we show more detailed differential gene expression between AKR1B10^{High} and AKR1B10^{Low} it will become clearer why GSEA did not show multiple GO term enrichment for AKR1B10^{Low}.

3.4.2.2 GSEA Hallmarks Enrichment

Another very useful tool of GSEA is the hallmark gene signature which includes 50 sets of approximately 4000 genes involved in a number of important and well-defined biological processes, thus they offer a much clearer picture of the differences between the cases being compared (173). The parameters for running GSEA on hallmark MSigDB (h.all.v6.2) were the same as those for GO Term enrichment. Only two hallmarks were significantly enriched in AKR1B10^{High} patients, reactive oxygen species (ROS) and MTORC1 signaling pathway (Figure 3.39) while there was no enrichment for AKR1B10^{Low}.

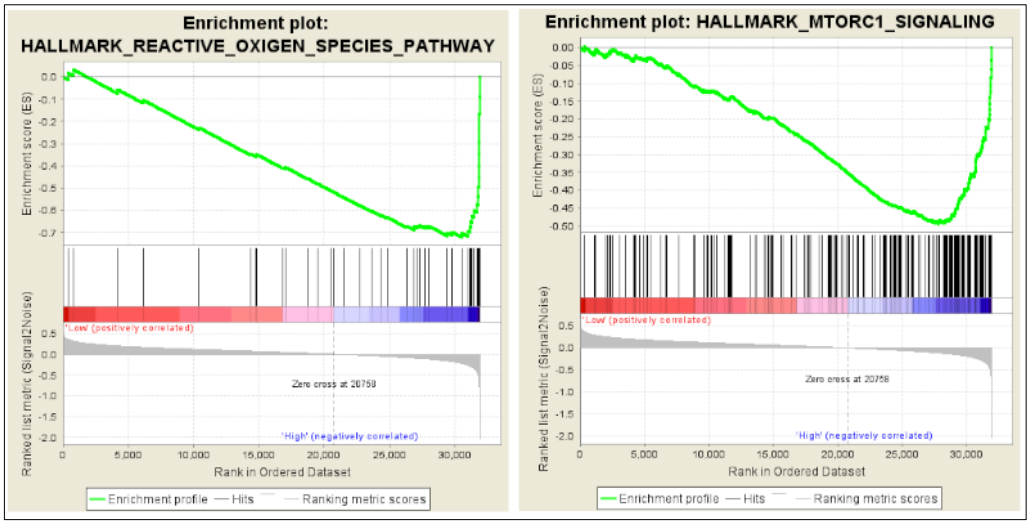

Figure 3.39 GSEA Hallmarks enriched in AKR1B10^{High} Samples.
The leading edge points down because of the order of the conditions fed into GSEA.

The first hallmark is composed of genes involved in NADPH metabolism and ROS response, as the name suggests, while the second hallmark is well-known to be involved in many aspects of tumor, especially metabolism and proliferation (174). Careful inspection of the enriched genes constituting the second hallmark shows that they are involved in amino acid metabolism, autophagy, proteasome activation, oxidative stress response and translation regulation. This is also in agreement with GO term enrichment shown above and in more detail in appendix A.

3.4.2.3 GSEA KEGG Pathways

The KEGG pathways MSigDB contains 186 curated gene sets representing the most important biological pathways according to their classification in the KEGG database. This is another compact way to represent results showing enriched biological pathways between two groups. There were 14 significantly enriched KEGG pathways in AKR1B10^{High} samples and none in AKR1B10^{Low} (FDR < 0.25). Figure 3.40 shows only 6 of these pathways while the whole list is shown in Appendix B. In addition to glutathione metabolism, one new pathway that shows significant enrichment is the pentose and glucuronate interconversion (PGI) which together with ascorbate and aldorate metabolism pathway (also enriched in AKR1B10^{High}) are important components of carbohydrate metabolism linking glycolysis and galactose metabolism with amino sugar and nucleotide sugars and producing esters and salts of glucuronic acid. The same pathway is also highly interconnected with hexosamine biosynthetic pathway and ascorbate and aldorate are essential components of antioxidation response in cells (128). This may also be connected to amino sugar and nucleotide sugar KEGG pathway which also showed significant enrichment (Figure 3.34). While this pathway has been shown to be dysregulated in a number of tumors, its real biological significance still remains obscure (102,175–177).

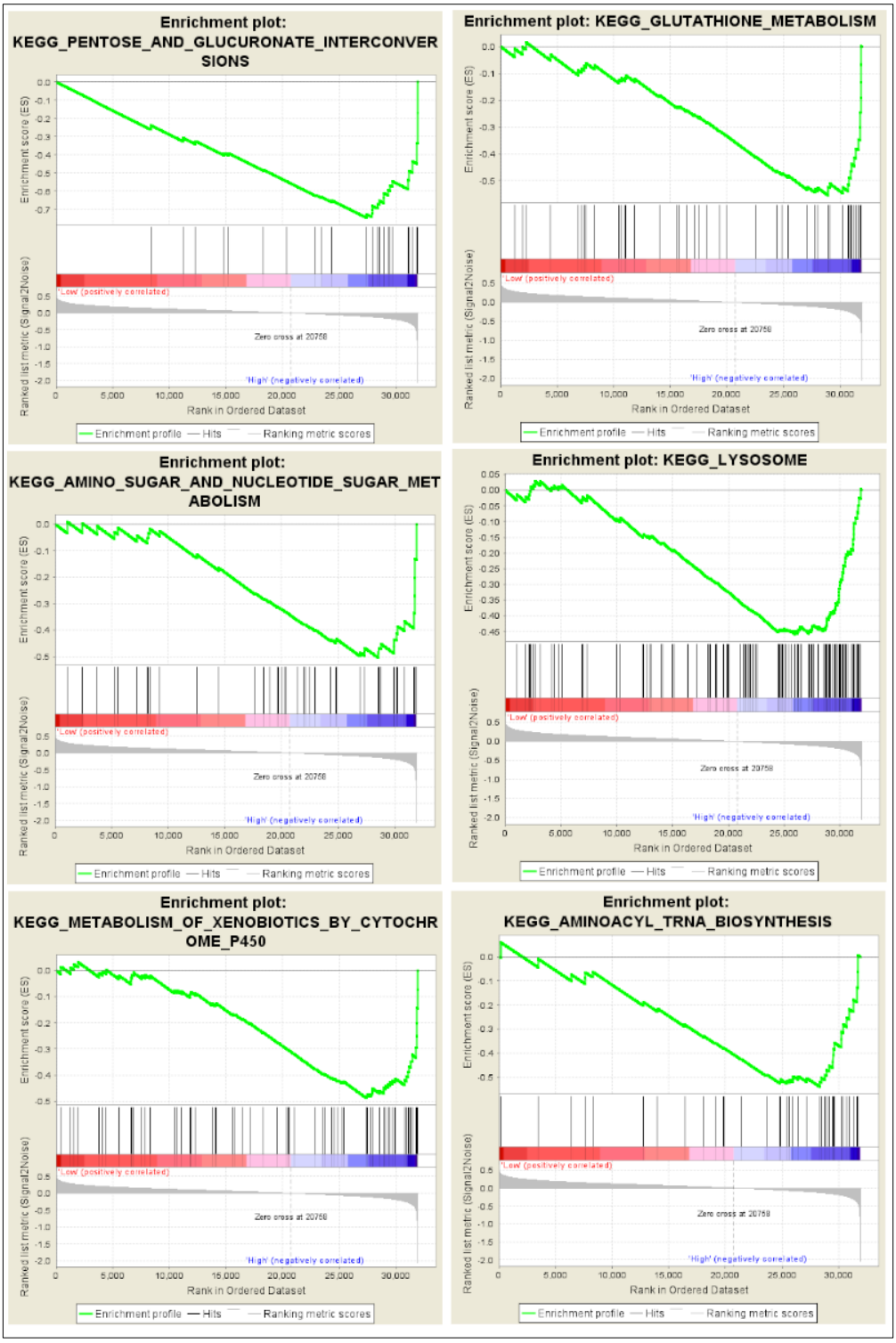

Figure 3.40 Six of GSEA KEGG Pathways Enriched in AKR1B10^{High} Samples.

The leading edge points down because of the order of the conditions fed into GSEA.

Figure 3.40 also shows enrichment of xenobiotic metabolism, which is mostly composed of AKR gene family members as well as UGTs. In addition, the Lysosome and amino acyl-tRNA pathways are involved in amino acid metabolism and protein translation. They are in agreement with mTORC1 hallmark enrichment shown in the previous section.

3.4.2.4 GSEA Reactome Pathways

The last MSigDB used in GSEA analysis was reactome which contains 1604 curated gene sets based on the canonical pathways defined in reactome database. This is especially important for our study because of its emphasis on biological reactions and it serves as a complementary approach to KEGG pathways and GO terms. There was only one significantly enriched reactome pathway in AKR1B10^{Low} (Figure 3.41) and 45 in AKR1B10^{High}, 6 of them shown in Figure 3.42 and the rest in Appendix C.

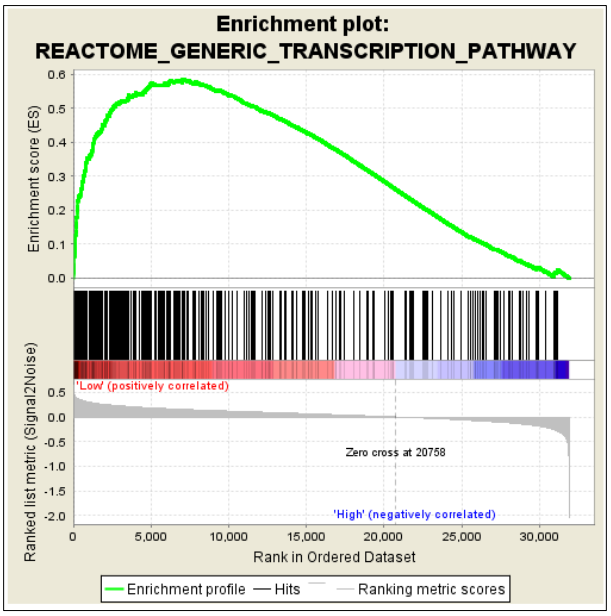

Figure 3.41 Reactome Pathway Enriched by GSEA in AKR1B10^{Low} Samples. This is in agreement with GSEA GO term enrichment shown in Figure 3.37.

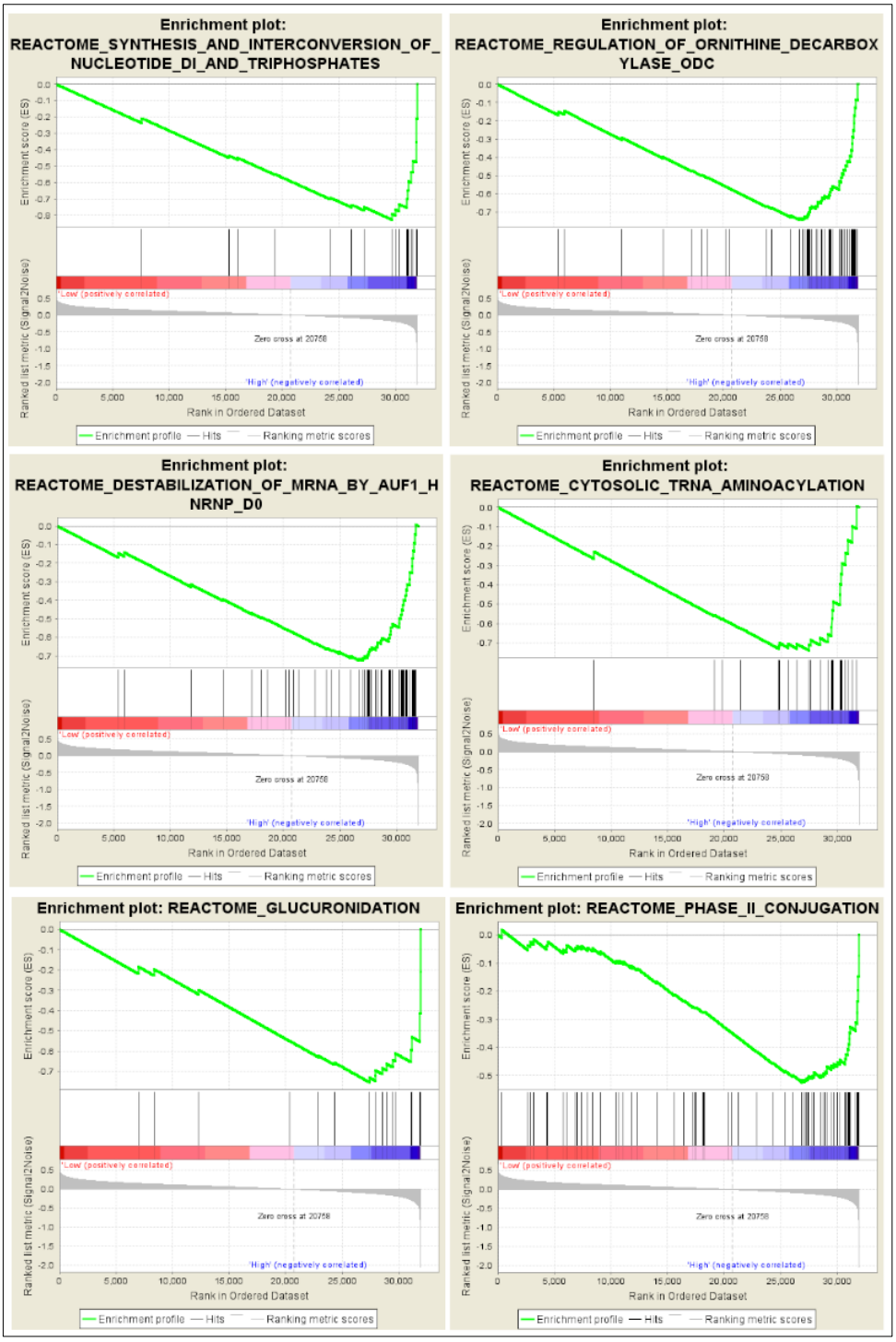

Figure 3.42 Six of 45 Reactome Pathways Enriched by GSEA in AKR1B10^{High} Samples.

The generic transcription pathway enriched in AKR1B10^{Low} samples includes hundreds of transcription factors, predominantly members of zinc-finger domain (ZNF) family, as well as some nuclear receptors. ZNFs constitute one of the largest families of transcription factors that are involved in many cellular processes such as DNA repair, cell migration and signal transduction, and especially transcriptional regulation. They also have been found to act as tumor suppressors or oncogenes depending on context (178).

The most significantly enriched reactome pathway in AKR1B10^{High} samples is involved in glutathione metabolism and oxidative stress response. The others are involved in amino acid metabolism, transcriptional and translation regulation by increased mRNAs stability and loading of amnio-acyl-tRNAs. In addition, cellular glucuronidation was found to be enriched in addition to phase II conjugation enzymes which, besides UGTs, include many members of amino acid lipid and glutathione conjugating enzymes. Other pathways shown in Appendix C are overall involved in ROS response, lipid and nucleotide biosynthesis and xenobiotic detoxification.

Observing so many processes related to drug detoxification, ROS response and amino acid metabolism, we took advantage of RPPA data available in TCGA. The RNA-Seq samples that were classified into AKR1B10^{High} and AKR1B10^{Low} were matched with RPPA and differential expression between the two conditions was performed using *limma*. Out of 184 RPPA samples, 91 were matched with our samples of interest, 39 for AKR1B10^{High} and 52 for AKR1B10^{Low}. Overall, out of 219 proteins with and without secondary modifications, 12 had significant difference ($p < 0.05$, $LFC > 0.2$) and their expression is shown graphically in figure 3.43.

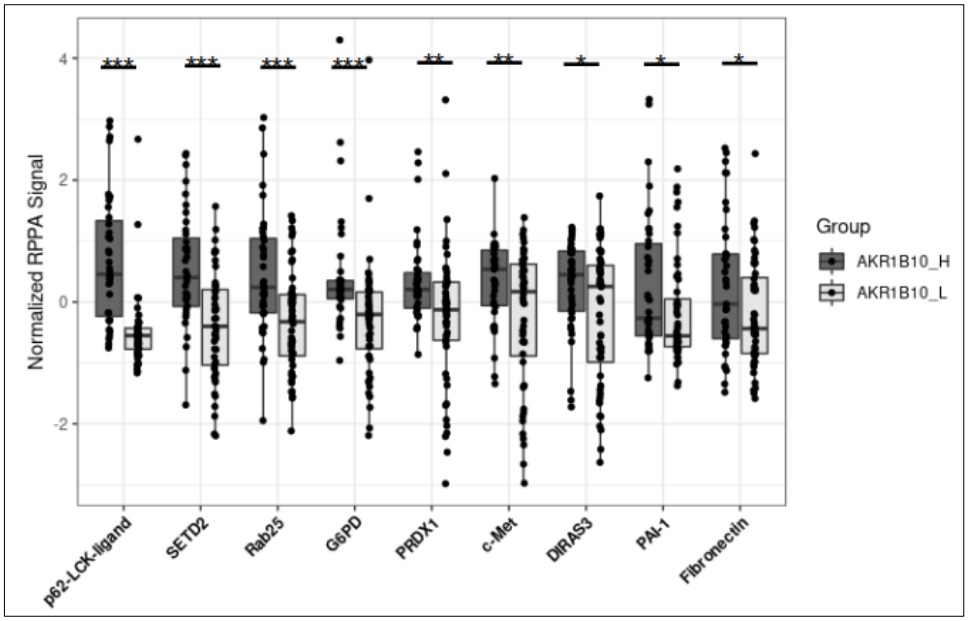

Figure 3.43 Significantly Differentially Expressed Proteins from RPPA Dataset.
After differential expression with *limma*, the expression values of these proteins were normalized and plotted. Significance levels: * p< 0.05, ** p< 0.01, *** p< 0.001 according to t-test.

RPPA confirms many GO, hallmarks, KEGG and Reactome pathways enriched by GSEA in AKR1B10^{High} samples (Figures 3.37, 3.39, 3.40, 3.42 and appendices A, B and C). In particular, mTORC1, ROS mitigation and amino acid metabolism related pathways such as proteasome degradation and lysosome were observed to be significantly altered. A significant increase in the expression of SQSTM1 (p62-LCK-ligand) was observed, which is known to play an important role in cytoprotection during oxidative stress by activating the antioxidative response through many pathways. In addition, it also acts as a mediator for autophagy which offers major survival advantage to tumor cells, especially in advanced stages (179). Phosphorylation of p62 was also found to channel glutamine towards the synthesis of glutathione and glucose to glucuronate pathway, giving liver cancer cells the ability to detoxify chemotherapeutic drugs and proliferate (180).

SETD2 is known to be involved in histone methylation and most of the studies have been concentrated on the association of its mutation with various tumors. Additionally, loss of expression has been associated with drug resistance but there is no study exploring its function when its expression levels are high (181). The mutation status of SETD2 was checked and it was found to be mutated in 4 samples of AKR1B10^{High} and 6 samples of AKR1B10^{Low}. Rab25 is a well-known oncogene which promotes cellular proliferation and survival. In addition, G6PD is the rate-limiting enzyme of the PPP reducing NADP into NDAPH while PRDX1 reduces various species of peroxides to water and alcohol. The other proteins such as c-MET, DIRAS3, Serpine1 (PAI-I) and fibronectin are also well-known to be implicated in various tumors.

The picture emerging when all the results are integrated together is that AKR1B10^{High} patients show increased expression of genes involved in NADPH synthesis, ROS response, xenobiotic detoxification, protein biosynthesis and autophagy. These are all important processes for cell survival and proliferation, and whether high expression of AKR1B10 has a central or marginal role is difficult to determine at this point. However, being able to utilize the expression of AKR1B10 as a biomarker helps in better understanding the presence of these processes which can then be studied in more detail to improve the clinical outcome of LIHC patients.

3.4.3 DGE Between AKR1B10^{High} and AKR1B10^{Low} Samples

In order to confirm some of the results of GSEA and potentially find others in addition to specifically evaluating the expression of metabolic enzymes introduced in Section 2.10, we carried out DGE analysis between AKR1B10^{High} and AKR1B10^{Low} samples using *DESeq2* package. Briefly, the same 180 samples used in GSEA were used for DGE by again dividing them into 2 categories, AKR1B10^{High} and AKR1B10^{Low} with 90 samples per group. A GLM with single variable for AKR1B10 status was used for the analysis. Overall, 2387 transcripts showed significant change between the two conditions (1LFC and FDR<0.05). However,

exploratory analysis on the nature of those transcripts showed that more than 40% were non-coding elements such as pseudogenes, lncRNAs, anti-sense transcripts etc. Moreover, the same elements constituted more than 50% of all the significant transcripts in AKR1B10^{High} samples. A summary of the most common elements is shown in Figure 3.44.

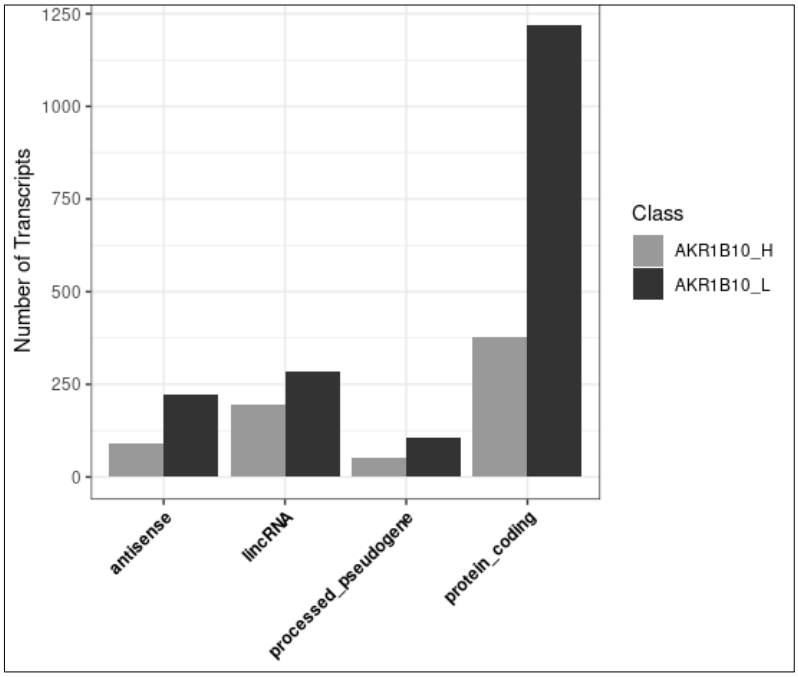

Figure 3.44 Transcript Class Distribution of the Significantly DEGs in AKR1B10^{High} and AKR1B10^{Low} Samples.

These four categories constitute 87% of the total significantly differentially expressed transcripts. The rest were removed for clarity.

As our focus was on coding genes, the non-coding elements were filtered out and the VST expression values of the remaining 1595 protein coding genes (377 overexpressed in AKR1B10^{High} and 1218 overexpressed in AKR1B10^{Low}) were used to construct a heatmap (Figure 3.45). The expression values were scaled by rows and the color was cut-off at 1 1LFC. AKR1B10^{High} samples formed an almost perfect

cluster, although not all the genes were uniformly upregulated, while AKR1B10^{Low} formed two separate clusters with no uniform expression (Figure 3.45).

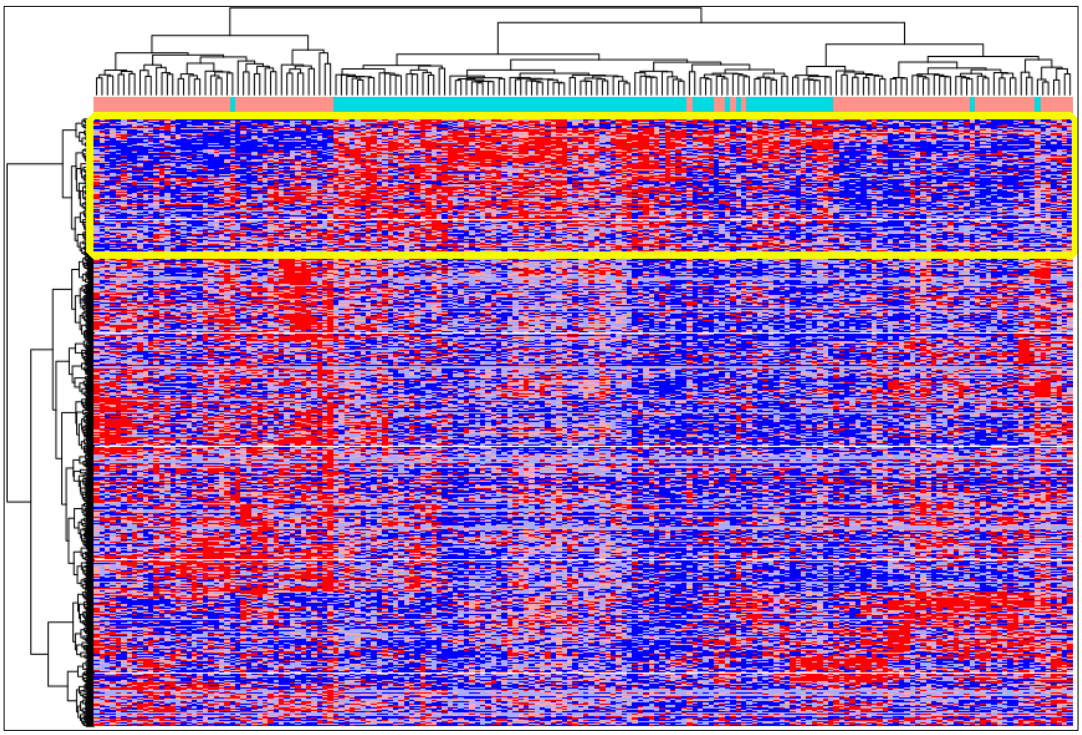

Figure 3.45 Heatmap of Significantly Differentially Expressed Genes.
Color Bar on top shows the two groups of patients: Cyan Represents AKR1B10^{High} and pink AKR1B10^{Low}. Clustering was done with Euclidean distance and Ward.D2 linkage and heatmap was generated with *pheatmap* package. Different gradients of blue and red color show genes with low and high expression, respectively. All AKR1B10^{High} genes form one cluster highlighted by the yellow box.

Figure 3.44 shows that AKR1B10^{Low} tumors were transcriptionally more heterogeneous, which explains the lack of GSEA enrichment observed in Section 3.4.2. We next carried out GO term enrichment with *clusterProfiler* package and obtained nearly 60 GO terms in AKR1B10^{High} samples. In order to remove redundancy between them, the GO terms were trimmed and reduced to 23 (Figure 3.46). In agreement with GSEA results, we obtained multiple GO terms related to

cellular glucuronidation, NADPH metabolism and cytokine/chemokine related pathways. However, as expected, they are not as comprehensive as GSEA. We also carried out GO term enrichment for AKR1B10^{Low} samples and obtained over 300 GO terms, however since their expression was not uniform throughout all the samples, they were not pursued any further.

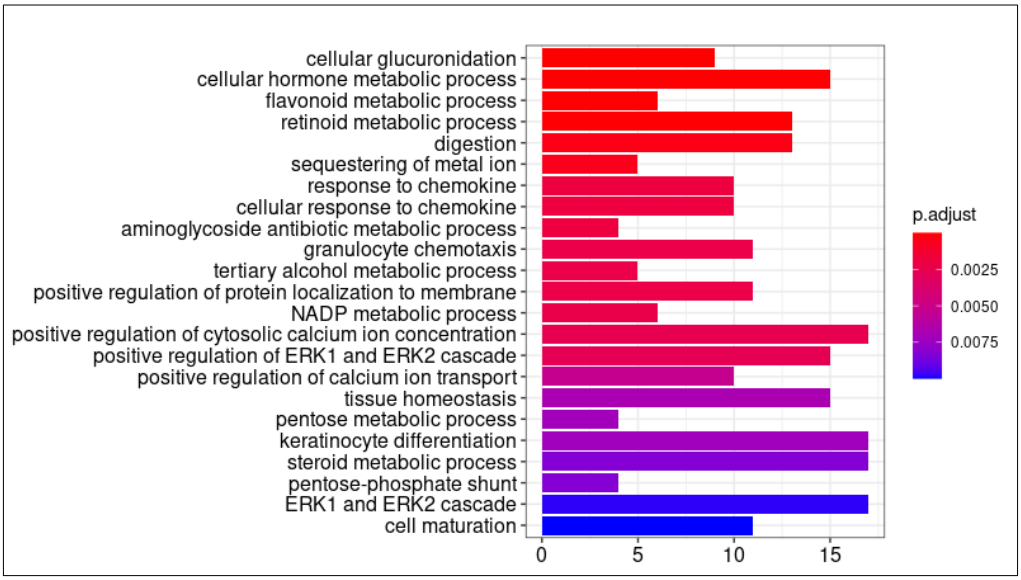

Figure 3.46 GO Term Enrichment in AKR1B10^{High} Samples.

These biological processes-related GO terms were obtained after filtering out redundant GO terms using simplify function.

Next, the 377 genes that were significantly upregulated AKR1B10^{High} samples in the highlighted cluster in Figure 3.45 were evaluated for the formation of gene interaction network in STRING server (https://string-db.org/). After gene symbols were uploaded and their interactions calculated in STRING, the network interaction scores were downloaded and analyzed in more detail in Cytoscape software version 8.2 and visualized (182). The genes formed a loose and complicated interaction network that was further analyzed with overlapping neighborhood expansion (ClusterONE) application. This software analyzes big networks and finds densely-connected subnetworks with stronger interactions. Using this approach, we obtained

two major densely-connected subnetworks. The network that was comprised almost exclusively of AKR, UGTs and NADPH-metabolizing enzymes is shown in Figure 3.47. Another major cluster was made of cytokines, interleukins and chemokines and it was connected with the cluster shown in Figure 3.47 via G6PD and epidermal growth factor (EGF). However, these genes were not consistently overexpressed in AKR1B10^{High} samples when the heatmap in Figure 3.45 was analyzed in detail and therefore not evaluated any further.

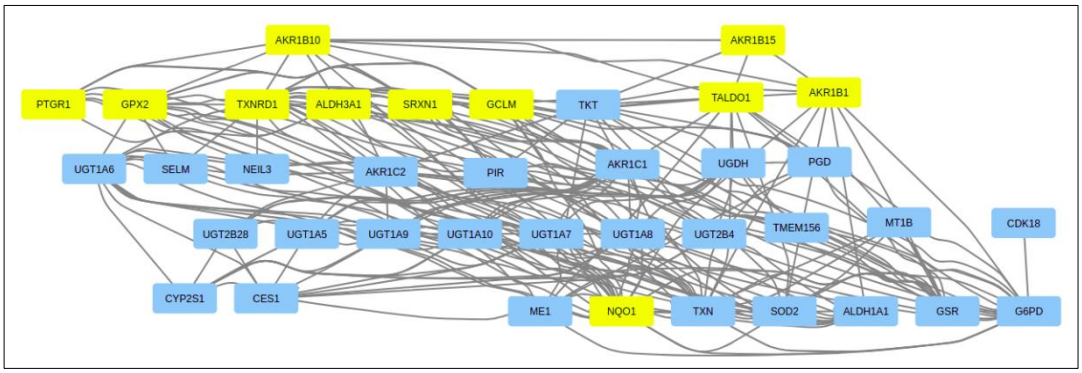

Figure 3.47 Gene Interaction Network in AKR1B10^{High} Samples.
Yellow boxes show genes that have direct interaction with AKR1B10. The network was created in the desktop application of Cytoscape.

For easier interpretation of the data, we evaluated the expression of the enzymes mentioned before mentioned in section 2.10 and found 184 of them to be significantly differentially expressed, 54 upregulated in AKR1B10^{High} and 130 upregulated in AKR1B10^{Low} samples (FDR < 0.05, 1LFC). Their expression through all samples is shown on heatmap in Figure 3.48. All the enzymes shown in the network in Figure 3.46 were clustered together on the upper half of the cluster highlighted by yellow box on the heatmap in Figure 3.48. These data highlight a coordinated upregulation of an entire gene network revolving around NADPH. Some of the enzymes such as G6PD, PDG and ME1 reduce $NADP^+$ into NADPH, while others such as AKRs and glutathione metabolism enzymes utilize it for their functions.

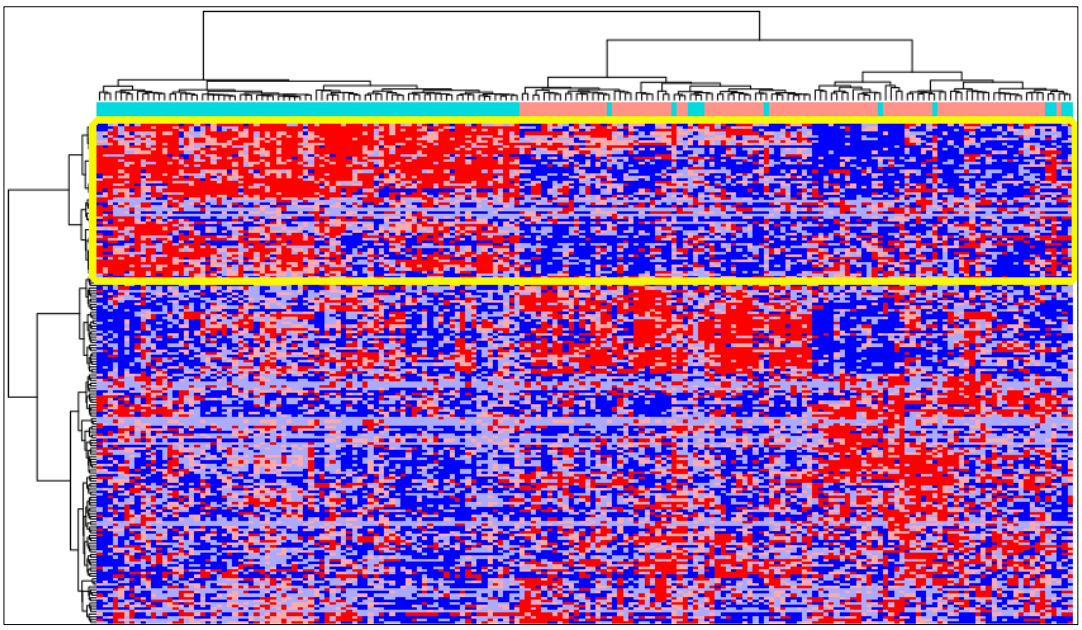

Figure 3.48 Heatmap of Significantly Differentially Expressed Enzymes.

Color Bar on top shows the two groups of patients: Cyan Represents AKR1B10^{High} and pink AKR1B10^{Low}. All AKR1B10^{High} genes form one cluster highlighted by the yellow box. Clustering was done with Euclidean distance and Ward.D2 linkage with *pheatmap* package.

In addition, another important enzyme enriched in the network is NEIL3, a glycosylase that functions in DNA excision repair by excising DNA bases damaged by oxidative stress. UGTs are phase II conjugating enzymes (170) that detoxify drugs. These enzymes may also function concordantly with AKRs, which are phase I conjugating enzymes, to detoxify xenobiotics or reactive molecules formed in the overstressed tumor environment. This may provide the AKR1B10^{High} class of tumors with a survival and proliferative advantage, leading to earlier patient death. The expression of these genes is shown in Figure 3.49.

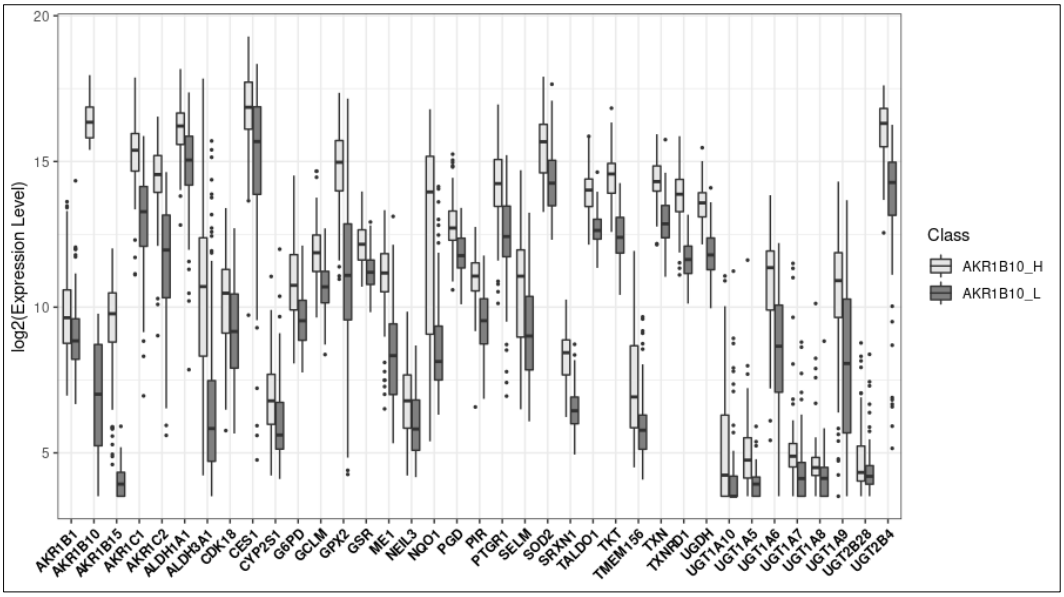

Figure 3.49 Expression of Network Genes in AKR1B10^{High} and AKR1B10^{Low}.
All of them showed statistically significant differential expression between conditions (Wald's test, $p < 0.05$)

3.4.4 Single Nucleotide and CNV Enrichment in AKR1B10^{High} and AKR1B10^{Low}

Comparative analysis for simple nucleotide mutations showed that TP53 was the most frequently mutated gene in 59 samples out of 180 (~33%). In general, most of simple nucleotide mutations were present in AKR1B10^{High} tumors, and the significantly enriched mutations ($p < 0.05$) are shown in the bar chart in Figure 3.50. Besides TP53 which is a well-known gene involved in genome quality control, some details about the functions and significance of mutations on the other genes will be given briefly. We also checked whether these mutations had a tendency to co-occur in the same patient, and for that again cBioportal server was used to analyze the mutations in AKR1B10^{High} and AKR1B10^{Low} patients separately. Figure 3.51 shows the oncoprints for both groups.

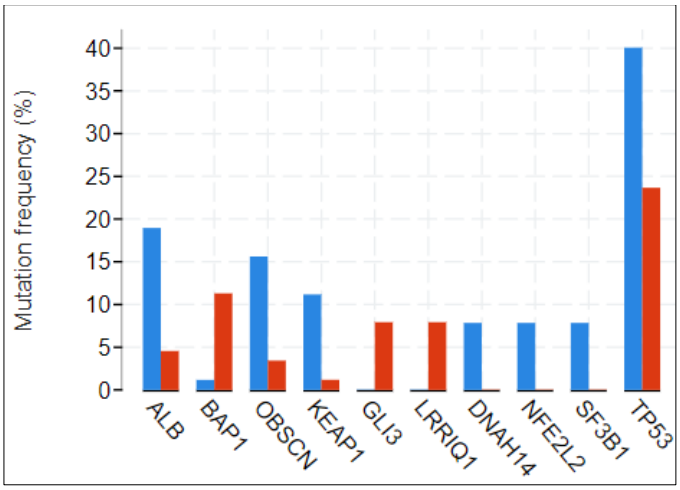

Figure 3.50 Genes Mutated in AKR1B10^{High} and AKR1B10^{Low} Samples.

Mutated genes are ordered according to their statistical significance. AKR1B10^{high} is represented by blue bar and AKR1B10^{low} by red bar. The figure was generated in cBioportal.

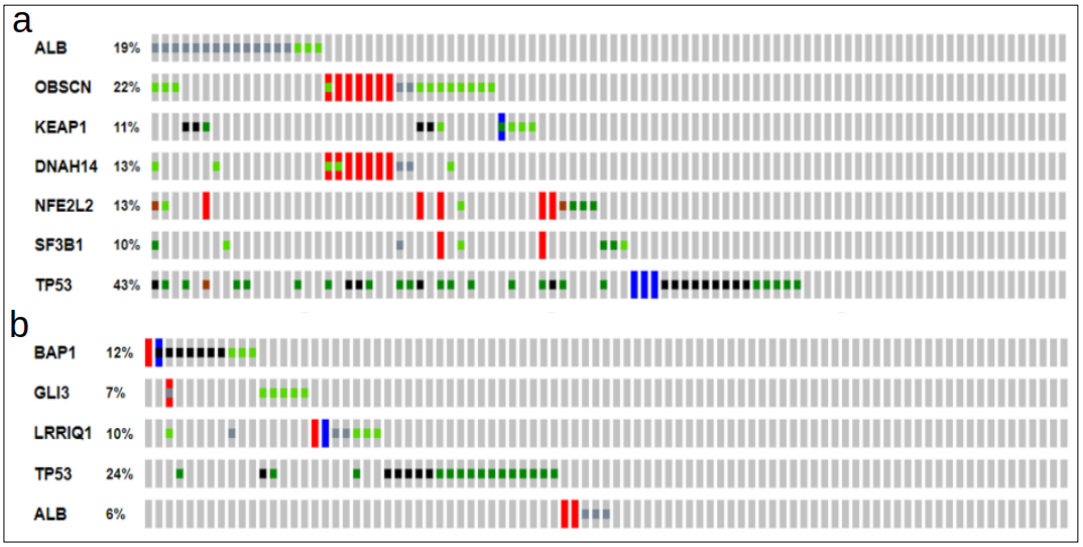

Figure 3.51 Onciprint Showing Significantly Mutated Genes.

(A) Genes mutated in AKR1B10^{High} and **(B)** genes mutated in AKR1B10^{Low}. Light blue, green and black show single-nucleotide missense and truncating mutations. Red and dark blue show amplifications and deletions, respectively.

Co-occurrence analysis showed that only OBSCN and DNAH14 had a tendency to co-occur within the same samples, but whether this is due to interactions between the two genes or was simply a confounding effect cannot be determined.

Serum albumin (ALB) is moderately mutated in liver tumors and studies have suggested a possible link to oxidative stress (183) which would fit with our model of AKR1B10 and PPP. BRCA1-associated protein-1 (BAP1) acts as a tumor suppressor via a deubiquitinase domain that regulates genes involved in DNA damage repair, transcription regulation and cell cycle. It forms complexes with BRCA1 and BARD1 proteins which enhance its ubiquitin ligase activity to control DNA damage and has been associated with higher risk of uveal melanoma (184).

Obscurins (OBSCN) are normally cytoskeletal proteins present in skeletal muscle with structural regulatory roles. However, this gene was recently found to be one of the most frequently mutated genes in colorectal and breast cancers, along with tumor suppressive functions in breast cancer (185). Tumor cells become resistant to apoptosis by either downregulating OBSCN expression, or by mutations which prevent them from carrying out their normal function (186). Kelch-like ECH-associated protein 1 (KEAP1) and the nuclear factor erythroid-2-related factor 2 (NRF2) signaling pathway were shown to have very important roles in xenobiotic and electrophile detoxification, as well as cellular response to oxidative stress. Mutation of KEAP in non-small cell lung cancer was shown to lead to a constitutive activation of NRF2, increasing resistance to chemotherapeutic drugs (187). These findings were confirmed by another study where mutations in KEAP1 were shown to lead to functional loss of the protein, and therefore increased activity of NRF2 in pulmonary papillary adenocarcinoma. As a result, tumors had increased drug resistance by increasing their antioxidation capacity (188). Recently, the SQSTM1-KEAP1-NRF2 axis has become a hot topic because of its involvement in cancer-promoting autophagy (179,180). Figure 3.43 shows that SQSTM1 was significantly upregulated at the protein level in AKR1B10^{High} tumors.

The other genes, despite statistical significance were enriched in only 10% of the samples. Briefly GLI3 is a member of Sonic-Hedgehoge (SHH) family and acts as a transcription factor, generally involved in developmental processes, while not much is known about LRRIQ1 except its high expression in the testis. Dynein Axonemal Heavy Chain 14 (DNAH14) is a member of motor protein families involved in motility and computational studies have shown that the mutations in this gene may act as tumor drivers in endometrial cancer (189). On the other hand, Nuclear Factor Erythroid 2 Like 2 (NFE2L2) is a transcription factor involved in antioxidant response by binding to genes with antioxidant response elements (ARE) and its mutation is associated with KEAP1 mutations (190). Finally, Splicing Factor 3b Subunit 1 (SF3B1) is subunit 1 of splicing factor 3b protein complex, which together with splicing factor 3a and 12S RNA form U2 small nuclear ribonucleoproteins complex involved in RNA splicing. Mutations in this gene were recently shown to be associated with poor prognosis in breast cancer and has been suggested as a breast tumor marker (191). Overall, proteins involved in oxidative stress response, in particular the KEAP1-NRF2, were found to have significantly more mutations in $AKR1B10^{High}$ tumor samples.

Finally, we evaluated differences in genomic rearrangements in the form of CNV between $AKR1B10^{High}$ and $AKR1B10^{Low}$ patients. We found that amplification was the main form of CNV that was significantly enriched in $AKR1B10^{High}$ tumors (FDR < 0.05). The procedure was similar to that explained in section 3.3.5. The total number of transcripts with CNV was 313, a significant portion of which were miRNAs and pseudogenes. More than 90% of these transcripts were enriched in $AKR1B10^{High}$ tumors (data not shown). Thus, both, mutation and CNV data suggest that $AKR1B0^{High}$ tumors have higher mutation rate and more genomic instability, and this may be as a result of more cellular proliferation and more oxidative stress, among other mechanisms that remain unexplored. These mutation data are also in agreement with the mutation data for 50 tumor samples matched with normal adjacent tissue described in Section 3.3.

3.4.5 Differential Methylation Analysis Between AKR1B10^{High} and AKR1B10^{Low}

We next carried out overall DMA analysis between AKR1B10^{High} and AKR1B10^{Low} groups and in contrast to the case when tumors were compared to normal samples, only approximately 9% of the total probes showed significant change (FDR < 0.05) and their numbers was almost equal. However, as the volcano plot is Figure 3.52 shows, there was stronger hypermethylation in the AKR1B10^{Low} samples, a state that can potentially explain the presence of more genomic instability in AKR1B10^{High} samples as well as the presence of GO terms related to methylation enzymes as shown in GSEA analysis (Figures 3.38 and 3.41).

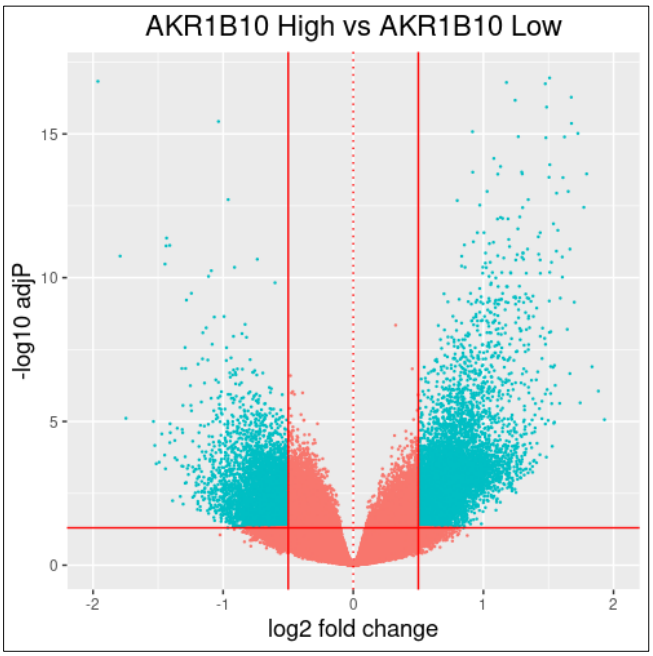

Figure 3.52 Volcano Plot showing DMA Analysis for AKR1B10^{High} and AKR1B10^{Low} Samples.

Probes hypermethylated in AKR1B10^{High} are represented on the right arm and those hypomethylated on the left arm. Y-axis shows the values of FRD in -log10 base. Vertical lines show the 0.5 LFC and horizontal line the FDR = 0.05. All probes passing these filters are colored in blue.

Next, we determined the methylation status of AKR1B10 and PPP genes on their TSS1500 in the same manner as we did with tumor and normal samples. We again found that AKR1B10 was hypomethylated in AKR1B10^{High} samples as compared to AKR1B1^{Low}, as expected, and correlation analysis between beta values and AKR1B10 expression in AKR1B10^{High} (but not AKR1B1^{Low}) showed significant correlation ($r = -0.34$, $p < 0.001$). We also found significant methylation differences for G6PD, TKT and TLDO1, but not for PGD, and these results, with the exception of TKT are in contradiction with the same analysis carried out on tumor versus normal section. However, as the boxplots in Figure 3.53 show, the distribution of beta values is very irregular, and we did not find correlation between the expression of any of the PPP genes and methylation values in any of the patient groups (data not shown).

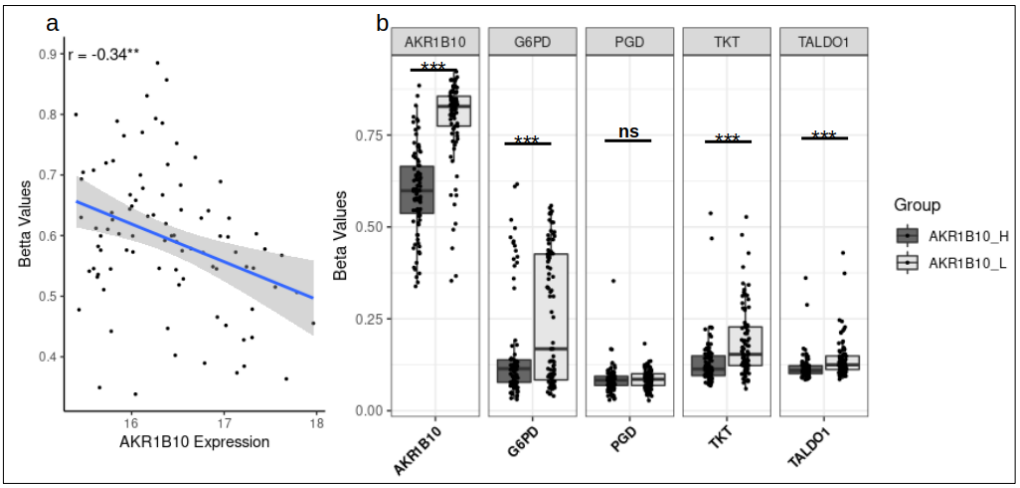

Figure 3.53 Methylation of AKR1B10 and PPP Genes in AKR1B10^{High} and AKR1B10^{Low} Samples.

(A) Pearson correlation of AKR1B10 expression and beta-values in AKR1B10^{High} tumors. Correlation coefficient is shown inside the plot **(B)** Probes used in this graph are: AKR1B10: cg11693019; G6PD: cg26799772; PGD: cg12796186; TKT: cg00223877 and TALDO1 cg08037817. Significance levels: * p< 0.05, ** p< 0.01, *** p< 0.001, ns not significant by t-test.

3.4.6 AKR1B10 and PPP in CCLE

In order to better understand molecular events and their regulation, tumors need to be modeled in cell culture, and the better the model mimics the real cases, the more realistic it becomes. For this purpose, publicly available transcriptional and metabolomics data of LIHC cell lines were analyzed. First, raw read counts for 25 liver tumor cell lines were downloaded from Broad Institute's CCLE database and normalized with the *DESeq2* package. The VST-normalized expression for AKR1B10, G6PD, TKT, TALDO1 and PGD were extracted and are shown in the bar plot in Figure 3.54.

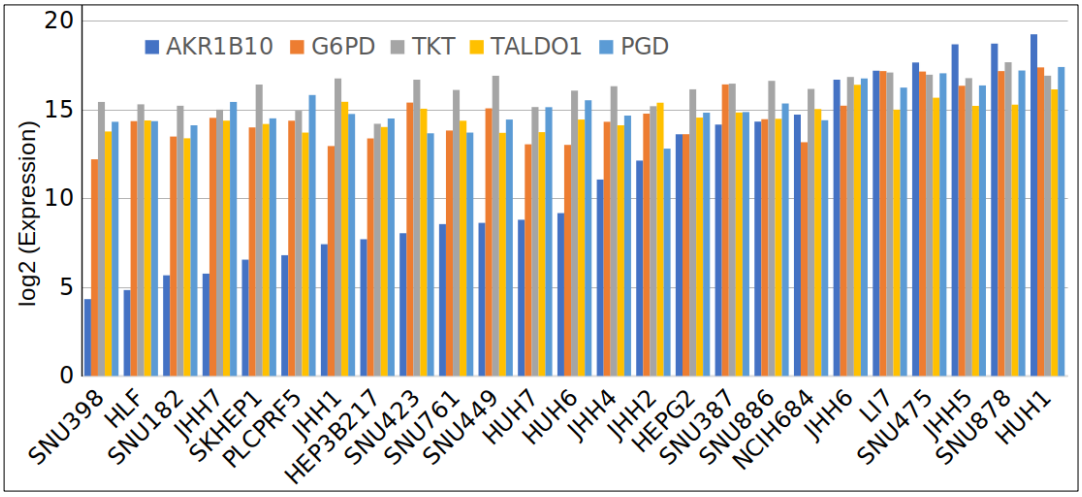

Figure 3.54 AKR1B10 and PPP Gene Expression in Liver Cancer Cell Lines.
Each cell is represented by 4 bars color-coded according to the legend on top. Cell lines are arranged according to AKR1B10 expression from the lowest to the highest.

CCLE expression data showed a wide range of AKR1B10 expression; however, the expression of the PPP enzymes did not change as dramatically, showing a profile similar to patient data. When the cells were classified according to AKR1B10 expression into high and low-expressing cells, two groups of 12 cell lines were formed with HUH6 falling in the middle. In order to mimic our model better, the highest- and lowest AKR1B10-expressing cells were chosen so to reflect significant

difference between PPP genes ($p < 0.05$, t-test). In this way, we created two groups separated into AKR1B10 and PPP low-expressing cells represented by SNU398, HLF, SNU182 and JHH7 and AKR1B10 and PPP high-expressing cells represented by SNU475, JHH5, SNU878 and HUH1.

Next, CCLE metabolomic data from CCLE database were utilized to find whether certain metabolites showed any difference between the two groups of cells mentioned above. Briefly, pre-normalized metabolite levels were downloaded from CCLE database. There were 225 metabolites that could be divided roughly into lipid and non-lipid, based on their chemical nature. First, we carried out a cluster analysis of all metabolites but were not able to identify any pattern matching the expression of AKR1B10 or PPP genes. Then metabolites were separated into lipid and non-lipid but again no clear cluster was observed (data not shown). Finally, in order to evaluate any the metabolite level differences between AKR1B10 high and low cell line groups, we carried out differential expression analysis between them using *limma* package. There was no significant change in metabolites between the groups for FDR < 0.05, except for 3 that showed significant p-value (Table 3.5). Since the metabolome data did not differentiate between oxidized and reduced forms of NADP which is at the center of this study, we did not pursue this line of experiments any further.

Table 3.5 Top 3 differentially expressed metabolites.

LFC stands for log-fold change, its positive values represent higher levels of metabolites on AKR1B10 high cells and vice-versa.

Metabolite	*LFC*	*p value*	*FDR*
Alpha Glycerophosphate	0.752146	0.0234	0.99
Butyrylcarnitine Isobutyrylcarnitine	-0.70525	0.0277	0.99
GABA	-0.78166	0.0305	0.99

3.4.7 CCLE Drug Sensitivity Analysis

In one of the landmark studies in the field of cancer therapeutics, nearly a thousand cancer cell lines were used to test the effects of multiple drugs, majority of them used in cancer chemotherapy. All the drugs were tested in various concentrations to determine the sensitivity and/or resistance of cancer cells (192–194). The data were stored in Cancer Therapeutic Response Portal (CTRP) and for this analysis the latest version of the server was used, as explained in section 2.6. If the correlation is positive, that means that higher expression of a gene is associated with greater resistance to a certain drug, while if the correlation is negative, higher gene expression is associated with more sensitivity towards the drug.

In total, 22 liver cancer cell lines were available in the CTRP database, and a search according to the method explained above yielded eight candidate drugs with high correlation coefficients and statistical significance. The drugs, their target and a short explanation are shown on Table 3.6. Drugs with the highest correlation with AKR1B10 expression are involved in NAD metabolism and ROS modulation.

Table 3.6 Drugs Correlated to AKR1B10 expression.

Compound	*Target*	*Explanation*	*Cor. Coef*
CAY10618	NAMPT	Inhibitor of nicotinamide phosphoribosyltransferase	0.74***
GMX-1778	NAMPT	Inhibitor of nicotinamide phosphoribosyltransferase	0.72***
Curcumin	Natural product	modulator of ROS and modulator of NF-kappa-B signaling	0.72***
BRD-M37545453	Natural product	modulator of ROS and modulator of NF-kappa-B signaling	0.70***
tigecycline	Unknown	Analog of tetracycline	0.942*
MST-312	TERT	Inhibitor of telomerase reverse transcriptase	0.66**
Ifosfamide	Unknown	DNA alkylator	0.86**
PI-103	MTOR, PIK3Cs	Inhibitor of DNA-PK, PI3K p110 delta, mTORC1, and catalytic subunits of PI3K	0.60**

Correlation plots for CAY10618, GMX-1778, Curcumin and PI-103 are shown in Figure 3.55. The plots show both the expression range of AKR1B10 in 20-22 cell lines and the AUC range of each drug. Since we could not find the AUC data for BRD-M37545453, the correlation shown on the table was taken from CTRP Server.

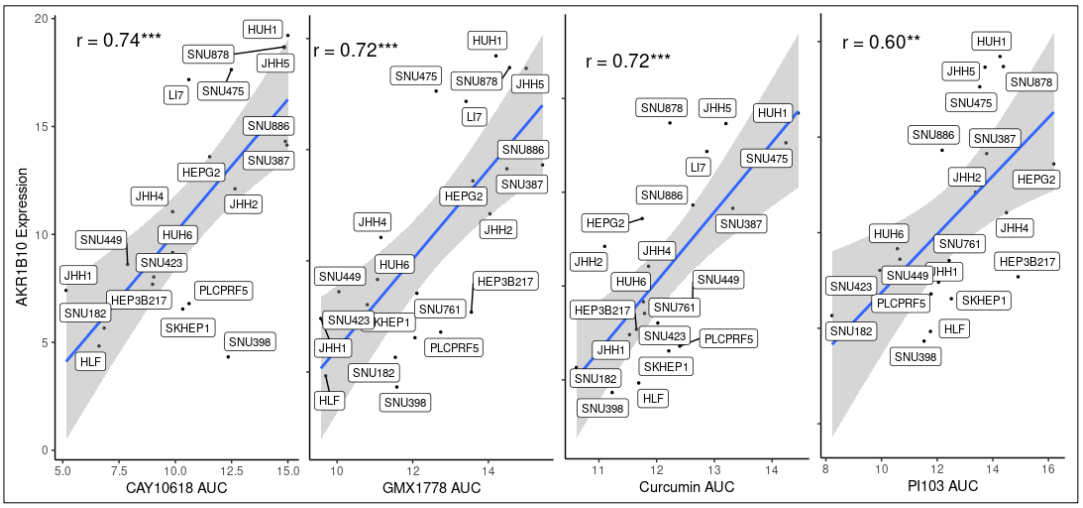

Figure 3.55 Correlation Plots Between AKR1B10 Expression and Selected Drugs. Y-axis shows the VST-expression values of AKR1B10 normalized by DESeq2 and x-axis the AUC values of each drug. Pearson correlation coefficients and their statistical significance are shown on the plots. The name of each cell line is shown on the plot.

The chemical structures of these 8 compounds were retrieved from CTRP database (http://portals.broadinstitute.org/ctrp/) and 7 of them are shown in Figure 3.56. The chemical structure of BRD-M37545453 completely overlapped with the structure of Curcumin. Judging the compounds by their structure, most of them are possible AKR1B10 substrates. CAY10618 and GMX1778 act by blocking NAMPT enzyme which is the rate-limiting enzyme in NAD^+ salvage pathway (195). The structural similarity of the rings of these compounds led us to speculate that these drugs may act as an NADP analogues. GMX1778 has an additional chloride moiety making it an extremely reactive group. The other compounds have many possible active groups which can be enzymatically reduced by AKR1B10, such as hydroxyl groups and chloromethyl radicals. However, to our knowledge none of these compounds has

been biochemically shown to be a substrate of AKR1B10 or any other AKR family member and further studies will need to be conducted on their possible interaction with AKR1B10.

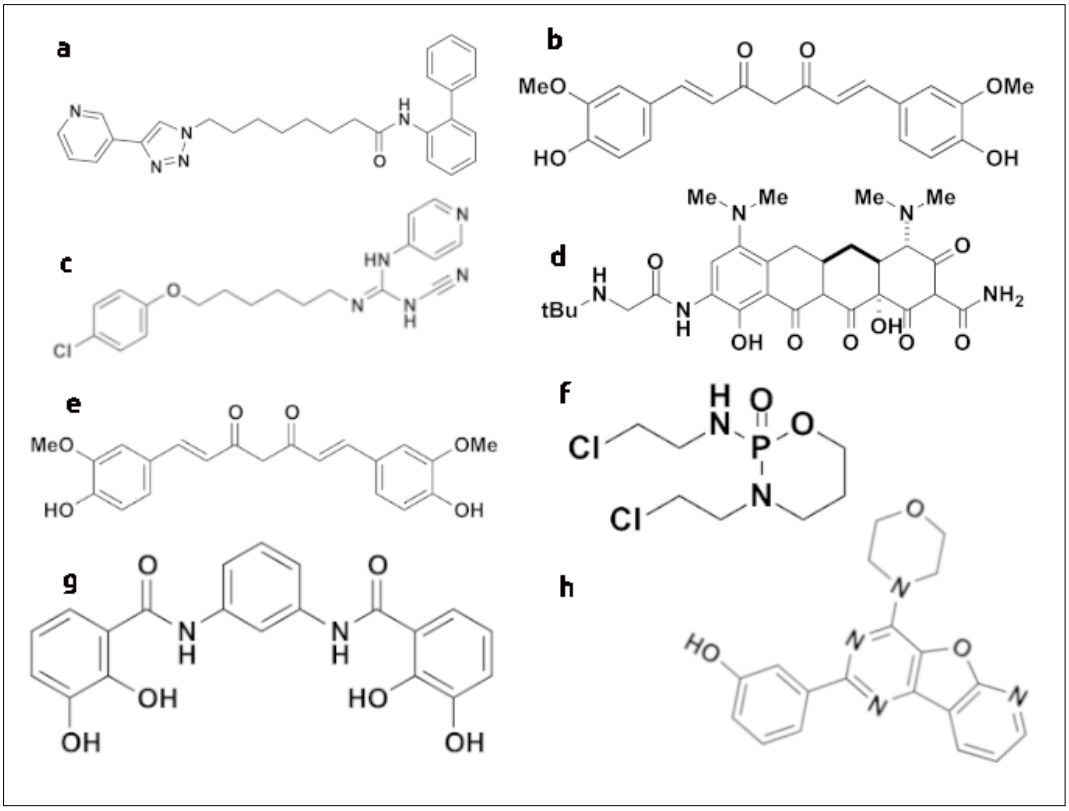

Figure 3.56 Structures of Drugs Correlated with AKR1B10 Expression.
(A) CAY10618, **(B)** GMX1778, **(C)** Curcumin, **(D)** MST312, **(E)** Tigecycline, **(F)** Ifosfamide, **(G)** PI103. All the structures shown here were downloaded from (http://portals.broadinstitute.org/ctrp/).

Among all the drugs, PI-103 is the only kinase inhibitor that blocks PI3K and mTOR. Studies have shown that it inhibits growth and proliferation of glioma cell lines as well as primary blast cells obtained from patients with leukemia (196,197). In addition, in the later study it was also shown that this compound induced significant apoptosis in leukemia stem cell population (198). Whether this is related to an

enriched mTOR pathway observed in AKR1B10^{High} patients needs to be further tested, but a positive correlation with AKR1B10 in an interesting observation.

3.4.8 Section Summary

In summary, in this section we showed that higher expression of AKR1B10 was associated with a median of more than 2 years shorter OS in LIHC patients. We then used two different approaches to understand the possible molecular mechanisms that can explain this difference: GSEA and DGE analysis. First, we used GSEA and found the patients in AKR1B10^{High} group showed enrichment of GO terms and biological pathways related to NADPH metabolism, both its synthesis by PPP, IDH1 and ME1, and its utilization by AKR enzymes and glutathione pathways.

In addition, we found the enrichment of another group of type-II conjugating enzymes, the glucuronyltransferases (UGTs) which we think may work in conjunction with AKRs which are type-I conjugating enzymes to detoxify both endogenous bioactive molecules as well as chemotherapeutic drugs. This group of enzymes together with a ROS response could be key to the survival and proliferation of AKR1B10^{High} expressing tumors in a highly hypoxic tumor environment. In addition, they may also be involved in drug detoxification and resistance during chemotherapy. This was also confirmed with DGE followed by gene network analysis. We found a complex network that could be broken down into two highly-connected subnetworks, the most enriched of which was composed of 38 highly connected genes almost all related to the enzymes involved in pathways involving NADPH and ROS (Figure 3.47). The second subnetwork was composed of cytokines, chemokines and growth factors which can also enhance tumor survival and proliferation.

Yet another important pathway we found to be enriched in AKR1B10^{High} samples is mTORC1 and amino acid metabolism, an observation that was also supported by RPPA data which indicate the activation of tumor-promoting autophagy. RPPA also

confirmed upregulation of G6PD and PRDX1 at the protein level, the first rate-limiting enzyme of PPP and the second essential for ROS response. AKR1B10^{Low} tumors on the other hand showed gene enrichment only for epigenetic-related pathways, which was supported by the DME analysis. Next, we evaluated the mutation status of AKR1B10^{High} and AKR1B10^{Low} groups and found enrichment of mutations in genes involved in oxidative stress affecting more than 50% of the AKR1B10^{High} samples. It should be emphasized that besides *TP53* which is a major tumor suppressor gene, mutation data showed an enrichment of KEAP1-NRF2 axis which together with an increase in SQSTM1 have been shown to play an important role in response to oxidative stress and tumor-promoting autophagy.

We also evaluated the methylation status of AKR1B10 and PPP genes on the TSS1500 probe. AKR1B10 and TKT, but none of the other genes appeared to be regulated by methylation. This is also in agreement with the methylation results we obtained by comparing tumors to normal samples in Section 3.3. In order to determine whether the metabolome was affected by AKR expression and to determine sensitivity to drugs, we utilized the CCLE and CTRP databases to determine the expression of AKR1B10 and PPP genes in LIHC cell lines as well as the accompanying metabolomics and drug sensitivity data. Currently there is no metabolome data for LIHC that is publicly available. Although the metabolomics analysis did not yield any useful results, we were able to utilize drug sensitivity data from large-scale experiments with drug compounds used in cancer chemotherapy. We were able to identify 8 drugs that correlated with AKR1B10 expression. Judging these compounds by their chemical structure and active groups, they appear to be ideal substrates for AKR1B10, however further biochemical and pharmacological tests are required to prove this assumption.

Overall, we have identified important biological pathways and molecular mechanisms involved in tumor-promotion in AKR1B10^{High} samples. AKR1B10 as an individual enzyme may be strongly involved in detoxification of bioactive molecules produced by highly proliferating cells and may be working in concert with UGTs. PPP on the other hand produces NADPH, which besides its use in oxidative

stress and almost all biosynthetic reactions, provides the necessary reducing power for AKR1B10 and other AKRs to carry out their enzymatic functions in the cell. These as well as tumor-promoting autophagy, cytokines and chemokines may all work in concert to promote tumor survival and proliferation, and thus earlier death in AKR1B10^{High} patients.

3.5 Multigene Risk Prognostic Model for LIHC

In the recent years, there has been an increasing number of studies utilizing machine learning (ML) approaches to stratify patients into different groups according to their overall or disease-free survival in various tumors such as glioblastoma (199), lung adenocarcinoma (200), ovarian cancer (201), bladder tumor (202) and liver (203,204). In addition, an earlier study took advantage of the power of deep learning to stratify liver patients into high and low risk groups (175). Although these models are very powerful in stratifying patients according to overall survival or disease recurrence, they suffer when it comes to clinical applications because they are based on the input of hundreds of genes that are then eliminated by dimensionality reduction techniques into a maximum of few dozen before used for cox regression analysis to calculate the risk level of each patient. Deep learning models on the other hand, despite their power because of deep neural network architectures and immense success in image classification (205), they are still performing not significantly better than simple regression models in clinical studies (206). In addition, they are "black box" models, i.e., we cannot really understand what is the weight, and thus the importance of each variable in the final classification model.

In this study, we used the genes from the network in Figure 3.46 to create a gene signature for stratification of liver patients into high and low-risk groups. We believe that the main advantage of this approach is its modular nature since we are concentrating on a handful of genes that are functionally related to each-other, in this way narrowing down the targets to one or few biological pathways related to NADPH metabolism.

3.5.1 Gene Selection for Cox Regression Analysis

We used the LIHC-TCGA dataset to train the multi-gene signature since it has all the available clinical data, and it also was the main dataset on which this study was based. For this purpose, we utilized the clinical data of 364 samples as mentioned at the beginning of section 3.4. The expression of 38 genes shown in figure 3.46 plus 2 other important enzymes of NADPH biosynthesis (IDH1 and PGLS) not included in the network because of their failure to reach 1LFC cut-off despite showing statistically significant expression difference between $AKR1B10^{High}$ and $AKR1B10^{Low}$ samples, was tested individually for correlation with patient overall survival by means of univariate cox proportional hazard model with *survival* package. Correlation coefficient and significance level for each gene are shown in Table 3.7, and based on these results, 17 genes showing significant correlation were used in the next steps of building a gene signature while the others were discarded. In order to find the correct gene combination for building the risk prognostic model, two approaches were used, step-wise systematic removal of non-significant genes and the more powerful LASSO regularization.

Table 3.7 Univariate Cox Regression Model Coefficients and Significance Values.

Gene	*Cox Coefficient*	*p-value*
AKR1B1	0.08253	0.206
AKR1B10	0.06026	0.0191
AKR1B15	0.1296	0.00104
AKR1C1	0.02926	0.568
AKR1C2	0.006866	0.873
ALDH1A1	-0.09454	0.0668
ALDH3A1	0.006783	0.794
CDK18	0.005456	0.923
CES1	-0.02257	0.58
CYP2S1	0.09841	0.109
G6PD	0.3208	2.16E-08

Table 3.7 (Cont'd)

GCLM	0.25891	0.00657
GPX2	0.017	0.614
GSR	0.3423	0.00112
IDH1	0.06197	0.593
ME1	0.13245	0.00944
MT1B	-0.1564	0.248
NEIL3	0.42119	3.41E-07
NQO1	0.07616	0.00827
PGD	0.30786	0.00164
PGLS	0.02106	0.84
PIR	0.05219	0.513
PTGR1	0.04196	0.442
SELM	0.05499	0.249
SOD2	-0.02048	0.783
SRXN1	0.27471	0.00124
TALDO1	0.19474	0.0428
TKT	0.21159	0.00252
TMEM156	0.05036	0.34
TXN	0.08263	0.372
TXNRD1	0.2446	0.00145
UGDH	0.16524	0.045
UGT1A10	0.15192	0.00521
UGT1A5	0.28119	0.00222
UGT1A6	0.07745	0.0923
UGT1A7	0.14985	0.0781
UGT1A8	0.255	0.0024
UGT1A9	0.008543	0.813
UGT2B28	0.1094	0.313
UGT2B4	-0.02923	0.496

3.5.2 Risk Prognostic Model based on Stepwise Gene Selection

In order to reduce the number of genes that could be fit to create a multi-gene cox regression model, we took advantage of *My.Stepwise* package which uses a mixture of forward- and backward propagation to select a list of the best genes for building a model which can then be modified by removing those with significance level of our choice. The whole LIHC-TCGA dataset was separated into training (n = 274, 75%) and test (n = 90, 25%) data sets. We started with the 17 candidate genes which showed individual correlation with patient survival in TCGA cohort and run the stepwise selection on the training dataset from which we obtained a list of 7 candidate genes, 4 of them significant ($p < 0.05$) which were retained, and 3 that were not significant were removed iteratively based on p-value.

The final model we used to calculate the risk score for TCGA and the independent cohorts was made of 4 genes, G6PD, NEIL3, AKR1B10 and AKR1B15 with cox coefficients 0.29483, 0.28709, -0.16976 and 0.22989, respectively ($p < 0.05$), which were used to calculate the risk score of every patient in the training set by the formula: 0.29483*(expression of G6PD) + 0.28709*(expression of NEIL3) - 0.16976*(expression of AKR1B10) + 0.22989*(expression of AKR1B15). We used *surv_cutpoint()* function for optimal separation of the groups into high- and low-risk groups with median survival of 12.3 and 69.6 months, respectively (HR = 5.88, log-rank $p < 0.0001$). The prediction accuracy for 1- 2- and 5-years was calculated by survivalROC() function (Figure 3.57).

The same coefficients were used to calculate the risk score of the test set (n = 90) which were separated into high-risk (n = 25) with median survival months 45.1 months and low-risk (n = 65) with median survival of 83.6 months (HR = 2.2, log-rank p = 0.033). These results are shown in Figure 3.58.

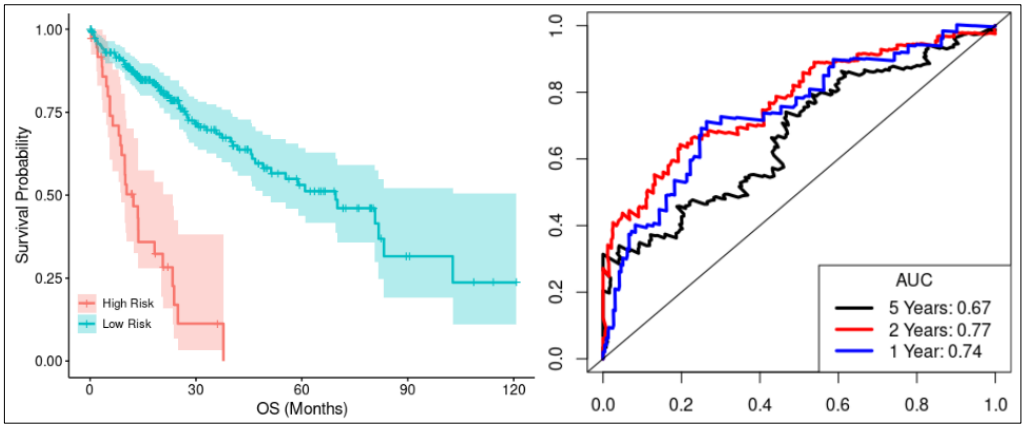

Figure 3.57 Patient Stratification in Training Dataset.

(Left panel) KM Plot showing patients in high-risk group (n = 39) and low-risk group (n = 235). (Right panel) AUROC curve with y-axis showing the true-positive rate while the x-axis showing the false-positive rate. The values of accuracy for 1-, 2- and 5-years are shown inside the plot.

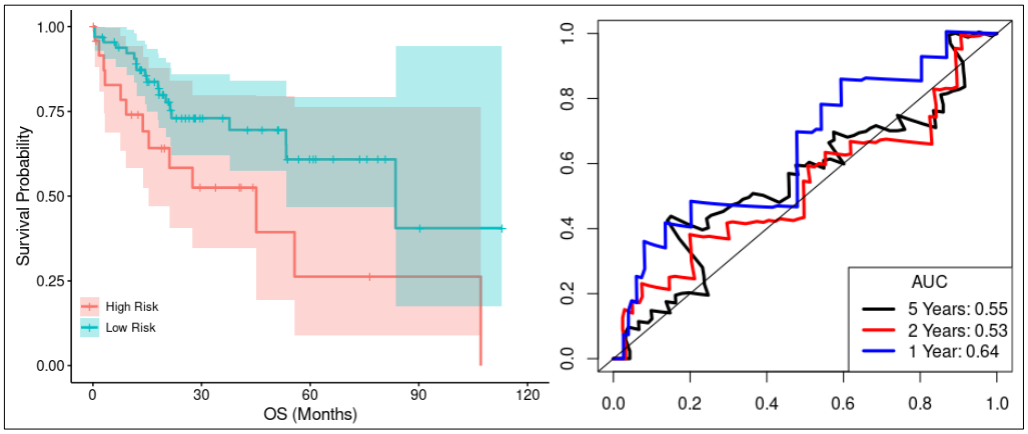

Figure 3.58 Patient Stratification in Testing Dataset.

(Left panel) KM Plot showing patients in high-risk group (n = 25) and low-risk group (n = 65). (Right panel) AUROC curve with y-axis showing the true-positive rate while the x-axis showing the false-positive rate. The values of accuracy for 1-, 2- and 5-years are shown inside the plot.

Finally, the same coefficients were used to calculate the risk score for the entire dataset and stratify all 364 patients into risk groups. Using the optimal risk cut-off, patients were stratified into high-risk (n = 45) with median survival of 12.3 months and low-risk (n = 319) with median survival of 69.6 months (HR = 5.2, log-rank p < 0.0001). The whole dataset was then separated into 4 groups as high-risk (n = 45) with median survival 12.3 months, intermediate-high (n = 114) median survival 55.4, intermediate-low (n = 114) median survival 58.9 and low-risk (n = 91), median survival 83.2 months. With this subgrouping, the hazard ratio between low-risk and high-risk groups increased to 6.73 (log-rank p < 0.0001). The KM plots for patient separation in 2 and 4 groups as well as the AUC curve are shown in Figure 3.59.

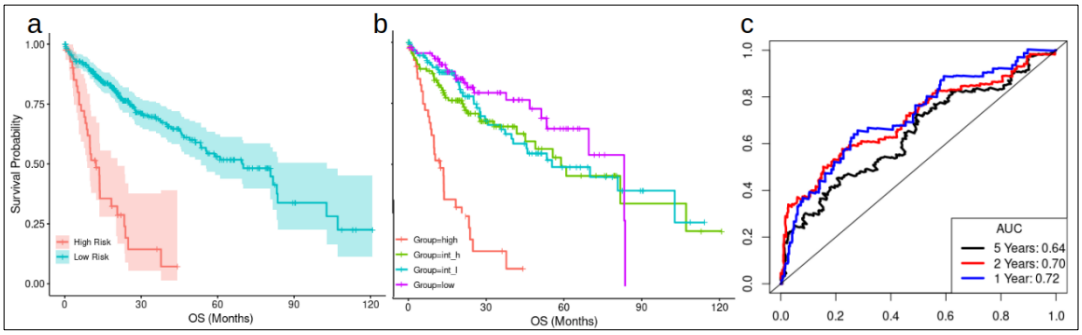

Figure 3.59 Patient Stratification in the Whole Dataset.

(A) Kaplan-Meier Plot showing patients in high-risk group (n = 45) and low-risk group (n = 319). **(B)** Re-stratification of patients into high- (n = 45), intermediate-high (int_h, n = 114), intermediate-low (int_l, n = 114) and low-risk groups (n = 91) (red, blue, green and magenta, respectively). Confidence intervals were removed for clarity **(C)** AUROC for the whole dataset, The y-axis shows the true-positive rate while the x-axis shows the false-positive rate. The values of accuracy for 1-, 2- and 5-years are shown inside the plot.

The accuracies in Figures 3.57-3.59 indicate model overfitting using the 4-gene signature. We observed a decrease in AUC in the test set and a significant recovery once the gene signature was refitted in the whole dataset since the test set makes up only 25% of the total data. Thus, although this gene signature was very good at stratifying the patients, it was not as accurate in predicting their OS.

The expression of genes used for generating the model is all significant between the groups (Figure 3.60).

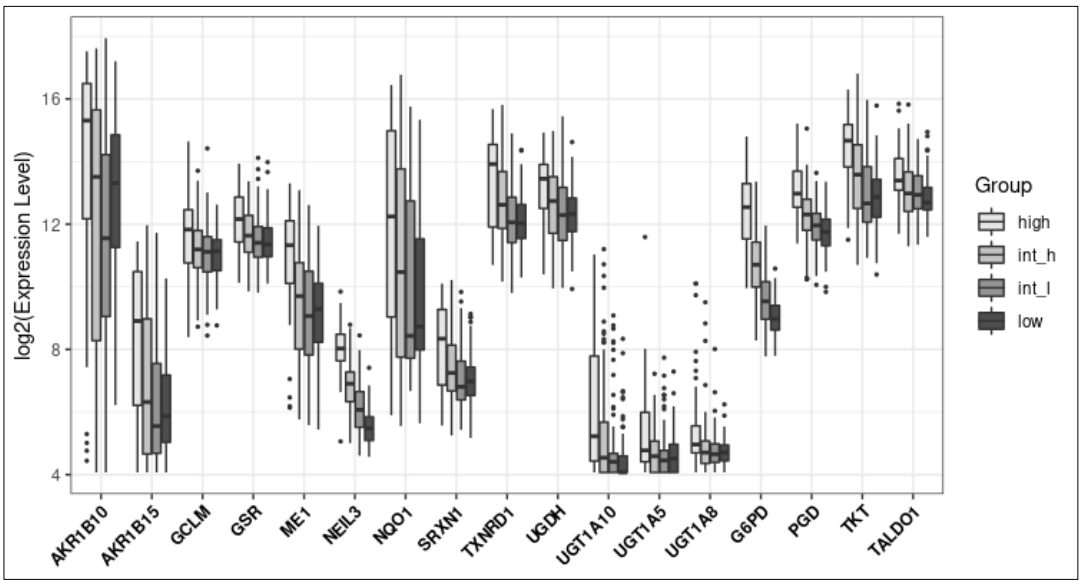

Figure 3.60 Expression of Genes Used for Signature.

The 4 conditions mentioned in figure 3.58 are color-coded. All the genes showed overall significant difference (One-Way Anova, $p < 0.05$).

The next step was fitting the model trained above into two independent datasets. First, we tried to fit the 4-gene model to GSE76427, however we failed to separate these patients into any significant group by using the coefficients calculated above so we tried a different approach by calculating the coefficients from GSE76427 and then separate the patients into prognostic risk groups. Although this is not how an ML approach would work, it is common practice in studies dealing with clinical data because of many factors, especially domain shift problems (207). It should be noted that none of the genes showed any significant correlation with patient survival in univariate as well as multivariate model. In addition, 16 patients with negative expression values for either AKR1B10 and/or AKR1B15 were removed in order to avoid biases since we were not able to process raw data as explained in section 2.1. Thus, the risk score for this dataset was calculated in the same way as TCGA training

set and patients were separated into high-risk group (n = 45) with median 6.29 years and low-risk (n = 46) group in which more than 50% of the patients survived during the time they were followed up as shown in Figure 3.61.

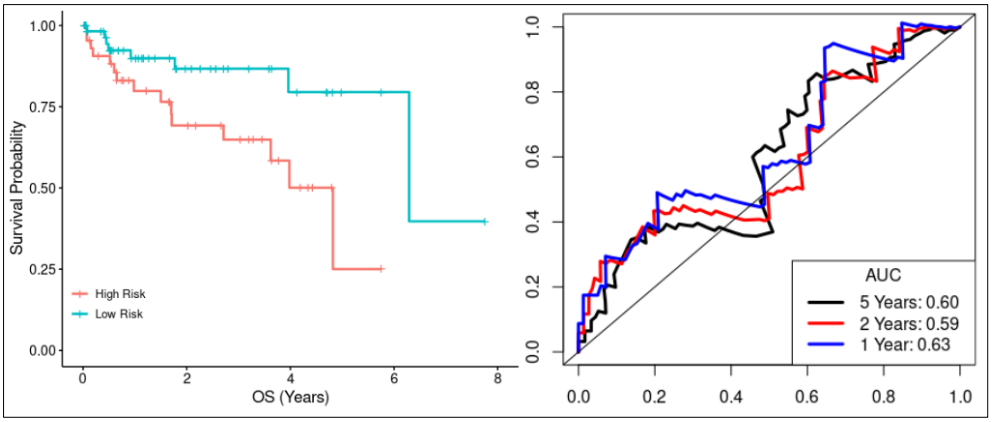

Figure 3.61 Patient Stratification in GSE76427 Dataset.

(Left panel) KM Plot showing patients in high-risk group (n = 45) and low-risk group (n = 46). The stratification was significant (HR = 3.6, log-rank p = 0.008(Right panel) the AUROC curve with y-axis showing the true-positive rate while the x-axis showing the false-positive rate. The values of accuracy for 1-, 2- and 5-years are shown inside the plot.

Finally, we tried to fit the same 4-gene model to the GSE14520 microarray, however there was no probe for AKR1B15 gene. As a result, only the coefficients of AKR1B10, G6PD and NEIL3 calculated from TCGA train dataset were multiplied with the expression of these 3 genes to calculate the risk score of each patient, who were then separated into two groups according to the optimal cut-off as high risk (n = 141) with median survival 54.8 months, and low-risk (n = 80) in which more than 50% of patients survived during the follow-up period. Their survival differences were significant (HR = 2.2, log-rank p = 0.0017). Both KM plot and the AUC curve are shown in Figure 3.62. For the time being, we will not be able to know whether the presence of AKR1B15 would have made the model better at stratifying GSE14520 patients and generalize well on independent datasets.

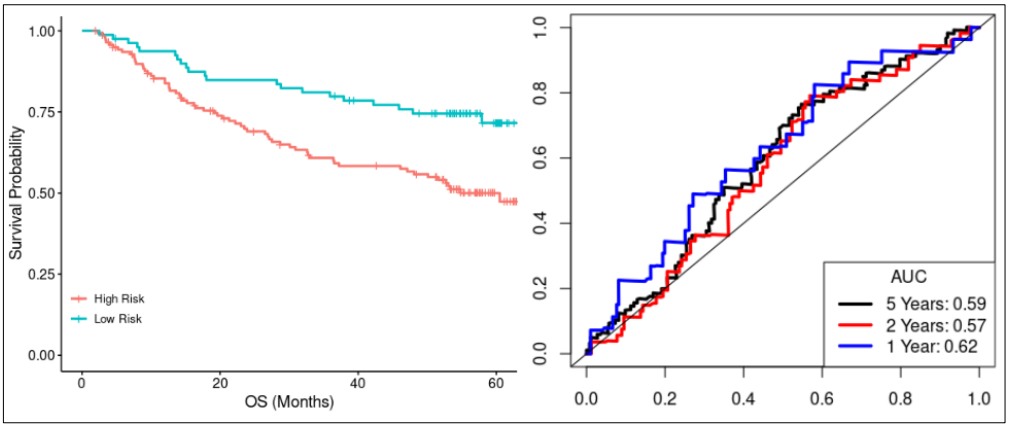

Figure 3.62 Patient Stratification in GSE14520 Dataset.

(Left panel) KM Plot showing patients in high-risk group (n = 141) and low-risk group (n = 80). The stratification was significant (HR = 2.2, log-rank p = 0.0017). (Right Panel) the AUROC curve with y-axis showing the true-positive rate while the x-axis showing the false-positive rate. The values of accuracy for 1-, 2- and 5-years are shown inside the plot.

We tried to train a model using the step-wise approach mentioned at the beginning of this section based only on G6PD, NEIL3 and AKR1B0 expression, but it did not show any significance. Thus, we utilized the power of LASSO regularization to build a 2-gene model as shown in the next section.

3.5.3 Risk Prognostic Model With LASSO

In the second approach, we used *glmnet* package to build cox regression model with LASSO regularization (123,208,209). The whole LIHC-TCGA dataset was again separated into training (n = 274, 75%) and test (n = 90, 25%) data sets by using the same seed number so that samples would be the same. Training data was used to find the smallest lambda value with various cross-validation folds and the best results were obtained with 10-fold cross validation, which was then used to select the genes with coefficients different from zero. Out of 17 genes, only two passed the filter, G6PD and NEIL3 while the others' coefficients were all reduced to zero as shown in Figure 3.63.

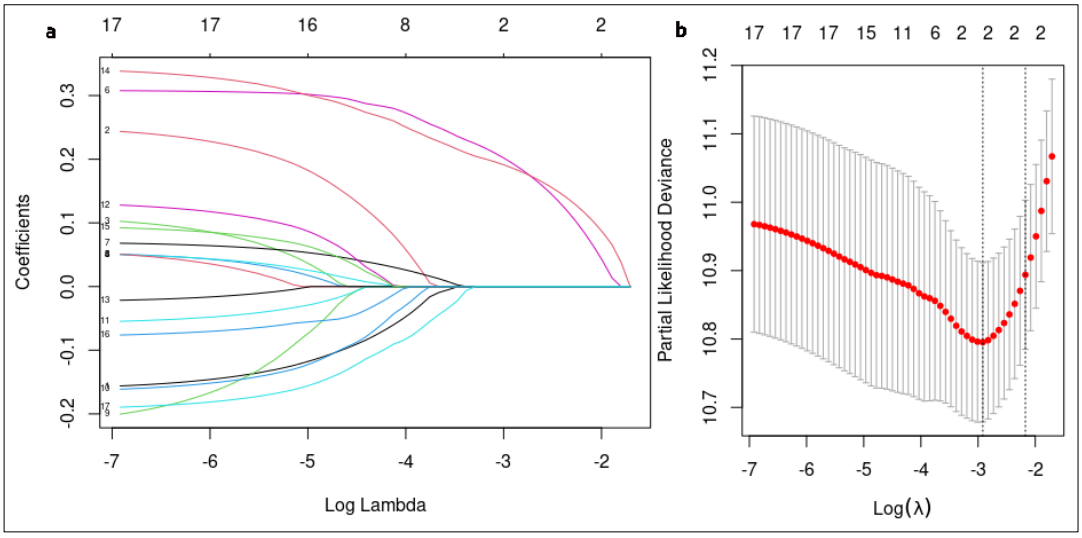

Figure 3.63 LASSO Regularization.

(A) Lambda coefficients plot, showing the decay in cox coefficients of each gene (y-axis) with increasing values of lambda (x-axis). **(B)** Partial likelihood deviance plot showing cox model fitting with 10-fold cross validation.

Using these two genes, a cox regression model was built on the training dataset. The coefficients of G6PD and NEIL3 were 0.24266 and 0.29085, respectively, and both were of high significance ($p < 0.01$). They were then used to calculate the risk score of every patient in the training set by the formula: 0.24266*(expression of G6PD) + 0.29085*(expression of NEIL3). Afterwards patients on training set were divided into high- and low-risk groups according to the optimal separation by using *surv_cutpoint()* function and their prediction accuracy for 1- 2- and 5-years was calculated by *survivalROC()* function as shown in Figure 3.64.

With these parameters, the separation between patients was highly significant (log-rank $p < 0.001$) and the hazard ratio for high-risk group was approximately 4.7. Median survival of high- and low-risk patients was 13.5 and 69.6 months, respectively. The same coefficients were also used to calculate the risk score for the test dataset and again patients were divided into high- and low-risk group (Figure 3.65).

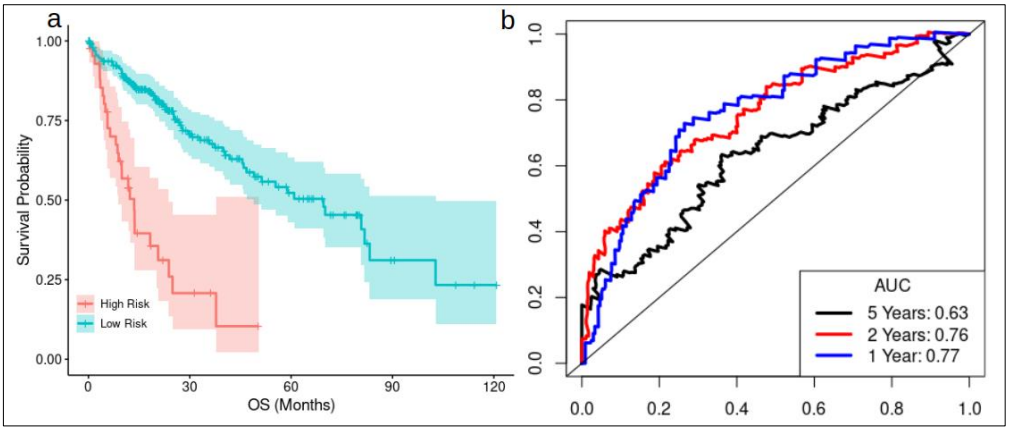

Figure 3.64 Patient Stratification in Training Dataset.

(A) Kaplan-Meier Plot showing patients in high-risk group (n = 44) and low-risk group (n = 230). **(B)** Area Under the Receiver Operating Characteristics (AUROC). The y-axis shows the true-positive rate while the x-axis shows the false-positive rate. The values of accuracy for 1-, 2- and 5-years are shown inside the plot.

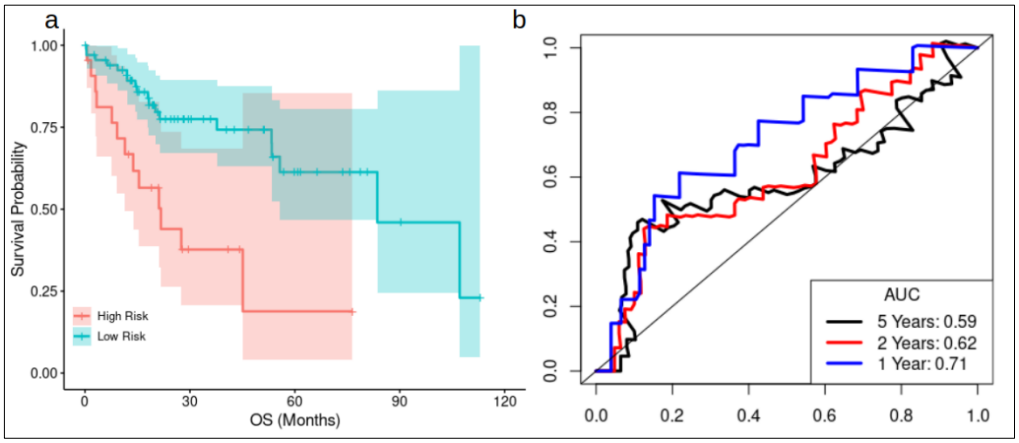

Figure 3.65 Patient Stratification in Test Dataset.

(A) Kaplan-Meier Plot showing patients in high-risk group (n = 22) and low-risk group (n = 68). **(B)** AUROC for the test dataset, The y-axis shows the true-positive rate while the x-axis shows the false-positive rate. The values of accuracy for 1-, 2- and 5-years are shown inside the plot.

Again, the separation between patients was significant (log-rank p < 0.01) and the hazard ratio for high-risk group was approximately 3.4. Median survival of high- and low-risk patients was 21.7 and 83.6 months, respectively.

Finally, the same coefficients were used to calculate the risk score for the patients in the entire dataset. Using optimal division cut-off, the median survival for high (n = 109) and low-risk (n = 255) patients was 21.7 and 70.1 months, respectively (HR = 2.8, log-rank p < 0.0001). In addition, the low-risk patients were grouped further into low risk (n = 89), intermediate low (n = 83) and intermediate high-risk (n = 83) and re-plotted in in KM Plot. This time the significance between high-risk group and low increased even further (HR = 3.2, log-rank p < 0.0001). Accordingly, median survival of low-risk patients was 69.6 months, that of intermediate-low was 83.6 months, intermediate-high was 55.7 months, and finally the median survival of high-risk patients was 21.7 months (Figure 3.66).

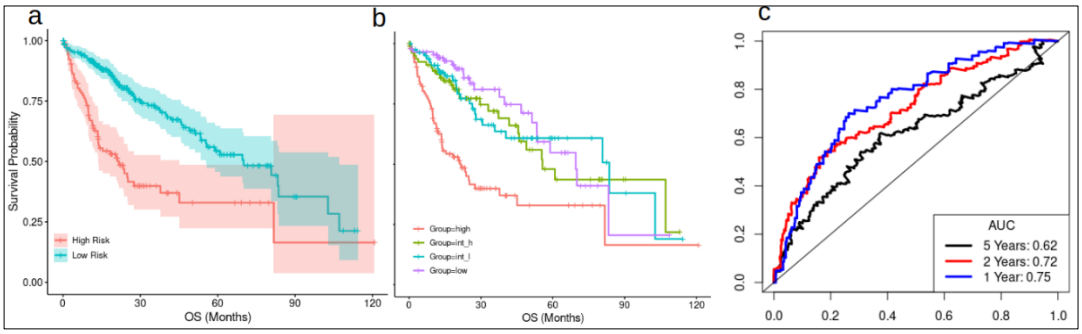

Figure 3.66 Patient Stratification in the Whole Dataset.

(A) Kaplan-Meier Plot showing patients in high-risk group (n = 109) and low-risk group (n = 255). **(B)** Re-stratification of patients into high- (n = 109), intermediate-high (int_h, n = 83), intermediate-low (int_l, n = 83) and low-risk groups (n = 89) (red, blue, green and magenta, respectively). Confidence intervals were removed for clarity. **(C)** AUROC for the whole dataset, The y-axis shows the true-positive rate while the x-axis shows the false-positive rate. The values of accuracy for 1-, 2- and 5-years are shown inside the plot.

As the KM plot shows in figure 3.66, the separation of patients into the KM plot is almost perfectly stepwise, with high-risk patients as an outlier and it gets better as the score gets lower. This 2-gene model also outperformed the stratification of patients in univariate cox analysis for each of the genes used separately. Model dependency analysis showed that high risk patients were associated with higher tumor grade, neoplasia grade and TP53 mutation (Appendix D). The expression of genes involved in model building was also checked and their expression patterns are similar to the Figure 3.59 but the differences between high and low-risk patients are slightly smaller, which could be the reason for the differences in HR (data not shown).

Different from the 4-gene signature, the differences in AUC between train, test and whole dataset did not change very significantly in this simpler model. This indicates that overfitting was not carried out; therefore, this signature has greater potential for generalization over other datasets as will be shown below.

Following the same procedure as the previous section, the correlation coefficients of the two genes were used to fit a similar cox regression model in two independent datasets. In GSE76427 it was again impossible to use obtain any significant results by using the coefficients generated by TCGA training data to build the model, most likely because of domain shift in the data (207). All the probes representing the G6PD and NEIL3 were checked for association with patient survival by univariate cox models but none of them showed significance. Then we followed the same procedure as section 3.5.2 by calculating the coefficients of the genes after building a cox regression model and then used them to calculate the risk score of each patient. The new score was able to stratify the patients into high- and low-risk groups with median survival of 4.82 and 6.29 years, respectively (HR = 3.04, log-rank p = 0.016) as shown in figure 3.67.

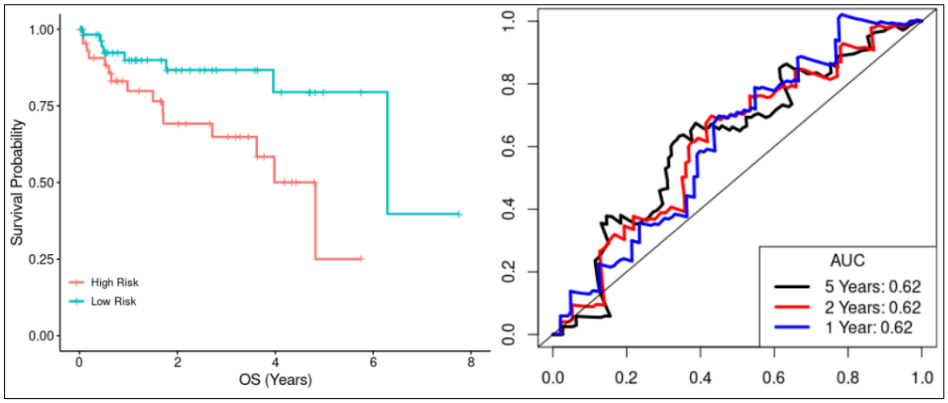

Figure 3.67 Patient Stratification in GSE76427 Dataset.

(Left panel) KM Plot showing patients in high-risk group (n = 48) and low-risk group (n = 67). The stratification was significant (HR = 3.04, log-rank p = 0.016). (Right panel) On the right, the AUROC curve with y-axis showing the true-positive rate while the x-axis showing the false-positive rate. The values of accuracy for 1-, 2- and 5-years are shown inside the plot.

We also used GSE14520 microarray dataset to test this model in the same way as in section 3.5.1 by using the coefficients calculated on the training dataset of TCGA. 221 patients were stratified into high (n = 149) and low-risk (n = 72) groups (HR = 2.75, log-rank p = 0.0004) by using the optimal stratification as well as by using the median of the risk score (HR = 1.70, log-rank p = 0.016). The KM Plot and AUC curve are shown in Figure 3.68.

Being able to fit the model into a completely independent dataset from a different platform and different population shows the potential of this gene signature to generalize as opposed to many other studies that calculate the gene signature by computing the correlation coefficients independently on every cohort used as external validation. The results obtained from validation of the model within the TCGA and also independent external validation especially on GSE14520 dataset shows the power of a very simple, 2-gene model to stratify patients at high-risk of dying from those who are not. More importantly, this model is based on a very

specific biological pathway which is related to NADPH synthesis, xenobiotic detoxification and oxidative stress.

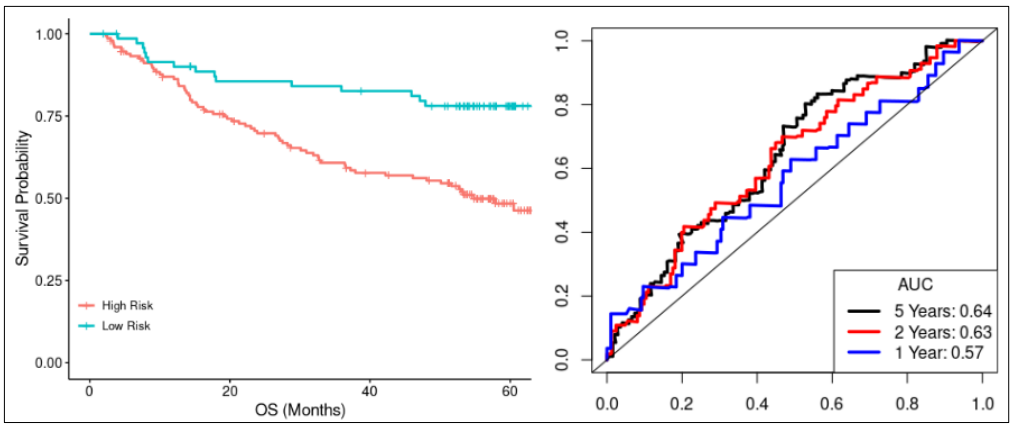

Figure 3.68 Patient Stratification in GSE14520 Dataset.

(Left Panel) KM Plot showing patients in high-risk group (n = 149) and low-risk group (n = 72). The stratification was significant (HR = 2.75, log-rank p = 0.0004). (Right panel) The AUROC curve with y-axis showing the true-positive rate while the x-axis showing the false-positive rate. The values of accuracy for 1-, 2- and 5-years are shown inside the plot.

3.5.4 DGE Between High and Low-Risk Patients Based on Two-Gene Signature

Studying and understanding the molecular mechanisms that cause such differences in patient survival would require a separate study, however, in order to get some understanding we carried out DGE of RNA sequencing data between the high- and low-risk groups for both gene models, and found that high-risk group, besides DNA replication and cell cycle, showed increased glycolysis and amino acid transport, and decreased fatty acid oxidation and amino acid metabolism. Surprisingly, this profile is very similar to the profiles we observed between normal and tumor samples in liver (Figure 3.23). Since detailed mechanisms causing these differences are not part of this thesis, only the expression of genes involved in glycolysis, PPP and hexose transport in the cell are shown in Appendix E. Patient stratification by 4-gene model

showed higher expression of PPP genes in high-risk group as compared to the low-risk one. Moreover, important glycolytic flux driving enzymes such as hexose transporter GLUT1 (SLC2A1), HK2/3, MCT-4 and PFKP/M were all overexpressed, generally at higher levels in 4-gene signature in which high-risk patients' survival was much lower than the 2-gene signature.

From these results, the importance of glycolysis in feeding PPP, among other pathways such as TCA is apparent. Additionally, the same set of genes involved in oxidative stress and xenobiotic detoxification (AKRs and UGTs) are all overexpressed in high-risk patients (data not shown). These results emphasize once again the idea that in the tumorigenic state, a complete reversal of normal metabolic functions of liver occur, such as increased glycolysis, decreased fatty acid and amino acid catabolism, and decreased urea cycle. Paying more attention to these in the future could lead to better management and even treatment of liver tumors maybe without any side effects as a result of drug toxicity which is very common in chemotherapy these days.

CHAPTER 4

CONCLUSIONS AND FURTHER STUDIES

Changes in tumor metabolism were proposed nearly 95 years ago by Otto Warburg who found that tumor cells, contrary to their normal counterparts, prefer to carry out fermentation even in the presence of oxygen, a phenomenon known as aerobic glycolysis or Warburg Effect. However, after the discovery of DNA and invention of genetic methods of cloning and sequencing, interest in metabolism fell and all the focus shifted in finding, tabulating and characterizing mutations associated with tumors. The initial hopes on the promise of massive sequencing technologies for detecting recurring mutations across cancers and developing drugs which would target the mutated forms of key regulatory proteins however are fading after many studies failed to extend the lives of patients despite their toxic side effects.

In a recent study, data from 127 clinical trials studying the effectiveness of 92 novel kinase blocker chemotherapeutic drugs approved by FDA for the period 2000-2016, showed that patient survival was extended by a mere median of 2.4 months (210). Actually, it can be safely claimed that most of the success in treating tumors to date can be attributed to early detection (211) and this realization has pushed the scientific community for more investment in developing better cancer detecting technologies aided by computer vision and machine learning, as well as awareness global campaigns for people to have regular check-ups. With the aging world population and increasing rates of diabetes, the incidence of tumor is estimated to surge and reach almost 30 million cases per year by 2040 (2).

Thus, the failure to find common mutations among tumors, or even within the same tumor on the same patient as a result of the huge cancer mutation landscape which gives tumors the capability to quickly develop drug resistance and recur in a more

aggressive form, has lead the research community to reconsider the ideas of Warburg and study cancer metabolism in hopes of better treatments. This was also aided by development in technology, which makes the study of metabolism and bioenergetics easier as compared to formal cumbersome biochemical techniques. Moreover, nowadays metabolism can be understood by using transcriptomics, while measuring the levels of the main metabolites in cell with modern techniques such as mass spectrometry has become routine.

4.1 Pan-Cancer Metabolism

Although bioenergetics is still an obscure field within the research community (212), it is believed that understanding metabolite flows within tumors could be the key in finding better treatments and management strategies for cancer in the future. The importance of metabolism and bioenergetics has started to be appreciated especially among evolutionary biologists who have recently challenged the widely-accepted model of "The RNA World". According to these studies, life did not begin with RNA, but it began in deep oceanic vents where complex reactions similar to the those of the TCA could take place. In addition, the most important event leading to the evolution of eukaryotic cell and all the complexity and diversity among higher life forms that we see today, including ourselves, was the result of endosymbiosis and formation of the mitochondria, which most likely paved the way for cellular compartmentalization leading to complex eukaryotic cell. These findings show that bioenergetics is a key factor in maintaining normal functioning of the cells and its disturbances could lead to diseases, including cancer. Moreover, we know that cells, in contrast to human-designed *complicated* systems are *complex* entities, a characteristic that gives them the ability to adopt, survive and evolve under different conditions, both favorable and unfavorable (213). The implications of these findings go beyond scientific curiosity as they have the potential to completely alter the current gene-centric view and lead to the next paradigm shift in molecular biology.

One of the main reasons I think metabolism should be understood well and in an unbiased way from molecular signaling pathways is that the concentration and flow of metabolites makes the survival of cells possible. Mutations are a normal part of life and their rate changes based on the food we eat, the environment we live in or even our genetic make-up. Therefore, there will always be potential for tumorigenesis. However, if the cellular context does not favor tumorigenic cells, they will never be selected and grow to kill the organism. I think the way tumor is currently understood is similar to separating evolution of life from its environment and laws of Physics, an idea that is completely unacceptable. Additionally, models like Knudson's "two-hit" hypothesis (3) or Vogelstein's "sequential model" (16) would work perfectly in a complicated system like a SpaceX rocket engine. However, they are logically and scientifically flawed to be portrayed as working models for a complex system like cells that interact with other cells, the whole organism and the environment in two-way dynamic fashion, and adjust themselves accordingly. Years of cell culture research have shown that cells have the ability to adopt to any treatment or modification that is not lethal.

In order to gain insights into the state of tumor metabolism, we took advantage of the publicly available transcriptomic data to study the expression of metabolic genes essential for cellular bioenergetics and biosynthesis. In contrast to many other studies that have aimed to understand tumor metabolism based on transcriptomics, we curated a panel of more than 600 genes involved in carbohydrate transport and metabolism through glycolysis, PPP, hexosamine and TCA, amino acid transport and metabolism through various pathways, fatty acid oxidation, synthesis and modification, anaplerotic reactions and electron transport chain (ETC), and showed their differential expression between tumor and matched normal samples on 20 tumor cohorts.

We found that overall tumors show increased carbohydrate transport into the cell, increased glycolysis, fatty acid and amino acid biosynthesis, along with decreased fatty acid beta oxidation, ketone body oxidation and amino acid catabolism. Since

eukaryotic cells can synthesize only a few amino acids through transamination reactions, they are auxotrophs and have to obtain them from the diet. Thus, amino acids are too valuable for tumor cells be used for energy because they are necessary for protein synthesis, making tumors reliant on amino acid import from the microenvironment, blood circulation or by activating autophagy. Reliance for short periods on amino acids can never be ruled out, however this is an exception and not the rule. Carbohydrates, on the other hand, are easy to obtain through the diet, especially currently, when most of the food and drinks are carbohydrate-based.

One major misconception about tumors is that all the glucose entering glycolysis is converted into lactic acid in order to oxidize NADH and thus ensure the continuation of glycolysis to produce the necessary ATP for other cellular functions. Lactic acid is then assumed to be excreted as waste product in the tumor microenvironment where its only function is to acidify the microenvironment and increase invasion. However, this assumption has two major flaws: first, although a certain fraction of glucose which, as our data shows can change from tumor to tumor, is indeed converted into lactate, the latter is not all excreted but can be metabolized by the cells. This has been recently shown in metabolite tracing experiments in mice (214,215). Second, downregulation of pyruvate dehydrogenase and pyruvate transporters in the mitochondria has been shown in many studies, including ours, but this does not mean that pyruvate is not utilized in TCA. This idea is flawed because tumors have been shown to produce more than 80% of ATP from the TCA more than 50 years ago (9).

The large quantities of pyruvate produced by elevated glycolysis more than count for the decreased expression of its transporters and pyruvate dehydrogenase. Moreover, pyruvate can enter the mitochondria through anaplerotic reactions and is converted into other TCA intermediates for replenishing oxaloacetate during periods of fatty acid biosynthesis, otherwise the TCA cycle would be shut down (216). Isotope-tracing experiments have shown that in tumor cells, most of the pyruvate is converted into citrate, which is then exported into the cytoplasm for fatty acid biosynthesis (217). These studies are in agreement with a model proposed in 2005

by Costello & Franklin (218) according to which, tumor cells have high rates of glycolysis to produce the necessary acetyl-CoA for fatty acid and cholesterol biosynthesis, a process that is still being overlooked. Here, we would add also that high rates of glycolysis are utilized for serine and NADPH biosynthesis, the first an essential amino acid for cell survival, and the second, the most important cofactor for lipid and nucleotide biosynthesis, as well as source of electrons to counter ROS in cells.

Our data suggest that in general, tumor cells have decreased levels of fatty acid oxidation, which is a major energy source in comparison to glucose or other molecules, including amino acids. Although there are studies showing that tumors utilize fatty acid oxidation for energy, they are almost all based on cell culture, which is a simplified representation of the more complex tumor tissue and its microenvironment. The reason why tumors have reduced fatty acid oxidation is connected with the anabolic state of the tumor. It would be biochemically and thermodynamically disadvantageous for the tumor to undergo a futile cycle of synthesizing and oxidizing its own fatty acids. The other reason could be the hypoxic environment and lack of oxygen, or mitochondrial defects. It is well known that fatty acid oxidation requires oxygen on which the electrons coming from carbon bond oxidation will be transferred for the cycle to continue.

Overall, our pan-cancer study shows that more attention should be paid to metabolism and better clinical approaches should be designed to control and manage blood sugar levels of tumor patients. This idea is also supported by clinical trials in which people undergoing intermittent fasting or kept under ketogenic diet show better response to tumor chemotherapy (219).

4.2 AKR1B10 and PPP

One of the key findings from the pan cancer metabolic enzyme assay was a major deregulation of NADPH metabolizing enzymes in many tumor types. AKR1B10 is

an NADPH utilizing enzyme that was remarkably upregulated in LIHC. G6PD, the rate limiting enzyme of the PPP was upregulated in nearly 50% of the tumors tested. The status of AKR1B10 as a liver tumor biomarker is under debate because different studies have shown opposing results regarding the association of its expression with tumor survival (Section 1.6.2). Our original hypothesis was to examine whether there existed a link between AKR1B10 (NADPH utilizing enzyme) expression and PPP (NADPH generating pathway) in AKR1B10^{High} tumor samples. Using various bioinformatic and computational approaches, we have shown a potential link between AKR1B10 and PPP in addition to other xenobiotic detoxification and oxidative stress mitigation pathways. Since PPP is one of the main NADPH synthesizing pathways in the cell, and most of these pathways utilize NADPH to carry out their normal function, enrichment of all them in a single group of patients with lower OS makes it a strong case.

Besides PPP and AKR1B10, upregulation of cellular glucuronidation and oxidative stress pathways in the same group of samples suggests the co-activation of cellular proliferation, mitigation of ROS and detoxification of bioactive compounds and/or chemotherapeutic drugs. PPP is a major branch of glycolysis that was found to be one of the most upregulated pathways in LIHC as compared to other tumors (Section 3.1.1). NADPH is the most important cofactor that is utilized in reducing cytotoxic active molecules, coping with high levels of oxidative stress and biosynthesis, especially fatty acid, steroids and nucleotides, the main macromolecules which in addition to amino acids are the basic necessities for cell proliferation. ROS, hypoxic environment and higher cellular proliferation rates form many bioactive molecules which if not neutralized, would wreak havoc and destroy the cells. In this setting, AKR1B10 as a phase I enzyme becomes extremely important in neutralizing many bioactive compounds as studies *in vitro* and *in vivo* have already shown (Section 1.6). UGTs are phase II enzyme that conjugate compounds with sugar moieties, making them easier for excretion (Section 3.4.2.1). AKR1B10 and UGTs have also been shown to be important in chemotherapeutic drug detoxification and tumor drug resistance. Thus, based on the data shown in this study, a whole pathway of

detoxification of toxic compounds and drugs can be envisioned, starting with phase I and being passed to phase II enzymes to be completely neutralized and excreted.

In addition, both GSEA and RPPA analysis also showed activation of autophagy and proteasome degradation pathways in AKR1B10^{High} patients, especially upregulation of SQSTM1 which has been shown to have important cytoprotective functions by activating antioxidative response and activate autophagy which offers major survival advantage to tumor cells, especially in advanced stages. All these pieces add up to create a cohesive model that makes possible higher proliferation and potentially drug resistance in AKR1B10^{High} patients. However, we cannot establish at this stage to what extent is the influence of high AKR1B10 expression in these patients.

In addition to other mechanisms proposed in the literature (90), here we show that the expression of AKR1B10 may also be controlled by methylation of TSS1500 region while the expression of PPP enzymes was not regulated in the same manner. We found a number of cell lines that can be used to better mimic the expression of AKR1B10 and PPP genes in laboratory experiments. Moreover, a number of drugs whose AUC values were positively correlated with AKR1B10 expression (representing resistance) were also identified for wet-lab experiments that are being carried out in our lab.

We also extended our findings to other datasets by training 2 Cox regression models, one including AKR1B10 and the other without it, and found that they can stratify LIHC patients better than any of these genes can do independently. We showed that these models have the potential to generalize over other independent cohorts. Although their accuracy is not as high as some gene signatures already published in the literature, their major advantage is on their nature as they are based on a few pathways with PPP at the center. However, these models should be tested in additional independent datasets in the future to establish their utility in stratifying LIHC patients into high and low-risk groups. DGE analysis between high and low-risk patients in LIHC showed that in addition to increased expression of AKR1B10, UGTs and ROS pathway enzymes, there was also increased levels of glycolysis, PPP

and amino acid transport, and decreased fatty acid oxidation and urea cycle. These results show that glycolysis and PPP are two underappreciated pathways in LIHC and their understanding could be key in better management and treatment of the second deadliest tumor in the world.

4.3 Limitations of the study and Future Directions

Despite the exhaustive nature of this study using multiple datasets and approaches to understand the changes in metabolic genes and the connection between AKR1B10 and PPP, the study has some limitations.

First, it is all based on *in silico* data, so having additional and better controlled liver tumors would add more strength to its results. An important step would be to use more independent datasets with LIHC patients and fit the cox model(s) in order to see whether they can be generalized on other populations and gene expression technologies, especially PCR which is widely used for gene expression quantification purposes. If this works, then the model(s) can be deployed for clinical use because it is fairly easy to quantify the expression of 2-4 genes from liver biopsy samples independent of any other information.

In case these results can be validated, the next important step would be to better study the molecular mechanisms and find ways to either target and block PPP which I believe to be the main driver of these pathways, which in turn is driven by glycolytic flux, or to find better clinical management protocols to reduce metabolite flux through these pathways. As it was mentioned briefly in Section 3.5.4 and shown in summarized form in Appendix E, glycolysis and PPP are at much higher levels in high-risk patients. Moreover, when the level differences between the high-risk groups defined by two models in TCG-LIHC are compared, the difference in gene expression between these pathways is higher in the 4-gene prognostic model where high-risk patients had lower survival.

Second, more rigorous and controlled experiments are necessary to better establish the role of AKR1B10 in the survival of these patients independently of other pathways mentioned in the previous section. Especially, understanding the role of AKR1B10 expression and function in the light of PPP is key in accepting, rejecting or modifying the results presented here. Also, it is important to establish whether AKR1B10, a phase I conjugating enzyme is working in concert with UGTs, phase II-conjugating enzymes in detoxifying bioactive compounds generated by oxidative stress or drugs used in chemotherapy. If there turns out to be a connection, this will be another important axis to study for improving chemotherapy outcome.

Third, as AKR1B10 is a drug metabolizing enzyme, experiments with some of the drug candidates identified in section 3.4.7 are necessary. Starting with simple in vitro experiments to show whether any of those drugs candidates are directly metabolized by AKR1B10 could be a be a beginning, and if some turn out to be metabolized by AKR1B10, more rigorous studies to understand their implication at cellular level will be necessary.

As a final shortcoming, we would like to highlight that this study involves the use of transcriptomics as a surrogate for metabolomics (Please see Section 1.8 for more details). This approach is not without drawbacks since we assume a strong correlation between the mRNA expression of enzymes and their enzymatic activities. However, the picture inside a cell is more complex and dynamic and the flow of metabolites is determined by many factors such as their concentration, enzyme availability, enzyme modulators and enzyme SNPs. We tried our best to eliminate the SNP factor because we checked the main enzymes for mutations and did not find any significant enrichment, however the other factors cannot be controlled. Thus, in the future, better controlled and isotope-tracing experiments possibly on live organisms such as mice coupled with mathematical modeling would provide very valuable data regarding metabolite flux in tumors as compared to normal tissues.

REFERENCES

1. Pardee A, Stein G. The Biology and Treatment of Cancer: Understanding Cancer. Second edi. Pardee A, Stein G, editors. John Wiley & Sons, Inc; 2009.
2. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2021;0(0):1–41.
3. Knudson AG. Cancer genetics. Am J Med Genet. 2002;111(1):96–102.
4. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell. 2018;173(2):291-304.e6.
5. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nat Med. 2004;10(8):789–99.
6. Hanahan D, Weinberg RA. The Hallmarks of Cancer. Cell. 2000;100:57–70.
7. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. Cell 2011;144(5):646–74.
8. Warburg, O., Wind, F., Negelein E. The metabolism of tumors in the body. J Gen Physiol. 1926;1(3653):519–30.
9. Pedersen PL. Tumor mitochondria and the bioenergetics of cancer cells. Prog Exp Tumor Res. 1978;22:190–274.
10. Warburg O. On the origin of cancer cells. Science (80-). 1956;123(3191):309–14.
11. Burk D, Schade AL. On Respiratory Imapairment of Cancer Cells. Science (80-). 1956;124:270–2.

12. Weinhouse S. On Respiratory Impairment in Cancer Cells. Science (80-). 1956;124:267–9.

13. Weinhouse S. The Warburg Hypothesis Fifty Years Later. Zeitschrift für Krebsforsch und Klin Onkol. 1976;87:115–26.

14. Sonnenschein C, Soto AM. Somatic mutation theory of carcinogenesis: Why it should be dropped and replaced. Mol Carcinog. 2000;29(4):205–11.

15. Marx J. Oncogenes reach a milestone. Science (80). 1994;266(5193):1942-4

16. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell. 1990;61(5):759–67.

17. Wishart DS. Is Cancer a Genetic Disease or a Metabolic Disease? EbioMedicine. 2015;2(6):478–9.

18. McLendon R, Friedman A, Bigner D, Van Meir EG, Brat DJ, Mastrogianakis GM, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455(7216):1061–8.

19. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin A V., et al. Signatures of mutational processes in human cancer. Nature. 2013;500(7463):415–21.

20. Zehir A, Benayed R, Shah RH, Syed A, Middha S, Kim HR, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat Med. 2017;23(6):703–13.

21. Fojo AT, Parkinson DR. Biologically targeted cancer therapy and marginal benefits: Are we making too much of too little or are we achieving too little by giving too much? Clin Cancer Res. 2010;16(24):5972–80.

22. Boroughs LK, Deberardinis RJ. Metabolic pathways promoting cancer cell survival and growth. Nat Cell Biol. 2015;17(4):351–9.

23. Mankoff DA, Eary JF, Link JM, Muzi M, Rajendran JG, Spence AM, et al. Tumor-specific positron emission tomography imaging in patients: [18F] fluorodeoxyglucose and beyond. Clin Cancer Res. 2007;13(12):3460–9.

24. Dang C V., Semenza GL. Oncogenic alterations of metabolism. Trends Biochem Sci. 1999;24(2):68–72.

25. Levine AJ, Puzio-Kuter AM. The control of the metabolic switch in cancers by oncogenes and tumor suppressor genes. Science (80-). 2010;330(6009):1340–4.

26. Vander Heiden MG, Cantley LC, Thompson CB. Understanding the warburg effect: The metabolic requirements of cell proliferation. Science (80-). 2009;324(5930):1029–33.

27. Gaude E, Frezza C. Tissue-specific and convergent metabolic transformation of cancer correlates with metastatic potential and patient survival. Nat Commun. 2016;7:1–9.

28. Wang Y, Xia Y, Lu Z. Metabolic features of cancer cells. Cancer Commun (Lond). 2018;38(1):65.

29. Harris AL. Development of cancer metabolism as a therapeutic target: new pathways, patient studies, stratification and combination therapy. Br J Cancer. 2020;122(1).

30. Altman BJ, Stine ZE, Dang C V. From Krebs to clinic: Glutamine metabolism to cancer therapy. Nat Rev Cancer. 2016;16(10):619–34.

31. Vettore L, Westbrook RL, Tennant DA. New aspects of amino acid metabolism in cancer. Br J Cancer. 2020;122(2):150–6.

32. Arruabarrena-Aristorena A, Zabala-Letona A, Carracedo A. Oil for the cancer engine: The cross-talk between oncogenic signaling and polyamine metabolism. Sci Adv. 2018;4(1):1–12.

33. Fan J, Ye J, Kamphorst JJ, Shlomi T, Thompson CB, Rabinowitz JD. Quantitative flux analysis reveals folate-dependent NADPH production. Nature. 2014;510(7504):298–302.

34. Zhou Z, Ibekwe E, Chornenkyy Y. Metabolic alterations in cancer cells and the emerging role of oncometabolites as drivers of neoplastic change. Antioxidants. 2018;7(1):1–18.

35. Trefely S, Lovell CD, Snyder NW, Wellen KE. Compartmentalised acyl-CoA metabolism and roles in chromatin regulation. Mol Metab. 2020;38(February):100941.

36. DeVita VT, Rosenberg SA, Lawrence TS. DeVita, Hellman, and Rosenberg's Cancer: Principles & Practice of Oncology. 11th editi. DeVita VT, Rosenberg SA, Lawrence TS, editors. LWW; 2018.

37. Thun M, Linet M, Cerhan J, Haiman C, Schottenfeld D. Cancer Epidemiology and Prevention. 4th ed. Oxford University Press; 2018.

38. Yang JD, Roberts LR. Hepatocellular carcinoma: A global view. Nat Rev Gastroenterol Hepatol. 2010;7(8):448–58.

39. Ally A, Balasundaram M, Carlsen R, Chuah E, Clarke A, Dhalla N, et al. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell. 2017;169(7):1327-1341.e23.

40. Marquardt JU, Andersen JB. Liver cancer oncogenomics: opportunities and dilemmas for clinical applications. Hepatic Oncol. 2015;2(1):79–93.

41. Moinzadeh P, Breuhahn K, Stützer H, Schirmacher P. Chromosome alterations in human hepatocellular carcinomas correlate with aetiology and histological grade - Results of an explorative CGH meta-analysis. Br J Cancer. 2005;92(5):935–41.

42. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. Science (80-). 2013;340(6127):1546–58.

43. Huang W, Li N, Lin Z, Huang G-B, Zong W, Zhou J, et al. Liver tumor detection and segmentation using kernel-based extreme learning machine. IEMBC. 2013;3662–5.

44. Das A, Acharya UR, Panda SS, Sabut S. Deep learning based liver cancer detection using watershed transform and Gaussian mixture model techniques. Cogn Syst Res. 2019;54:165–75.

45. Zucman-Rossi J, Villanueva A, Nault JC, Llovet JM. Genetic Landscape and Biomarkers of Hepatocellular Carcinoma. Gastroenterology. 2015;149(5):1226-1239.e4.

46. Bruix J, Takayama T, Mazzaferro V, Chau GY, Yang J, Kudo M, et al. Adjuvant sorafenib for hepatocellular carcinoma after resection or ablation (STORM): A phase 3, randomised, double-blind, placebo-controlled trial. Lancet Oncol. 2015;16(13):1344–54.

47. Utsunomiya T, Shimada M, Kudo M, Ichida T, Matsui O, Izumi N, et al. Nationwide study of 4741 patients with non-B non-C hepatocellular carcinoma with special reference to the therapeutic impact. Ann Surg. 2014;259(2):336–45.

48. Muto Y, MORIWAKI H, SAITO A. Prevention of second primary tumors by an acyclic retinoid in patients with hepatocellular carcinoma. N Engl J Med. 1999;340(13):1046–7.

49. Abou-Alfa GK, Venook AP. The impact of new data in the treatment of advanced hepatocellular carcinoma. Curr Oncol Rep. 2008;10(3):199–205.

50. Llovet JM, Ricci S, Mazzaferro V, Hilgard P, Gane E, Blanc J-F, et al. Sorafenib in Advanced Hepatocellular Carcinoma. N Engl J Med. 2008;359(4):378–90.

51. Cheng AL, Kang YK, Lin DY, Park JW, Kudo M, Qin S, et al. Sunitinib versus sorafenib in advanced hepatocellular cancer: Results of a randomized phase III trial. J Clin Oncol. 2013;31(32):4067–75.

52. Briggs A, Daniele B, Dick K, Evans TRJ, Galle PR, Hubner RA, et al. Covariate-adjusted analysis of the Phase 3 REFLECT study of lenvatinib versus sorafenib in the treatment of unresectable hepatocellular carcinoma. Br J Cancer. 2020;122(12):1754–9.

53. Toh HC, Chen PJ, Carr BI, Knox JJ, Gill S, Ansell P, et al. Phase 2 trial of linifanib (ABT-869) in patients with unresectable or metastatic hepatocellular carcinoma. Cancer. 2013;119(2):380–7.

54. He L, Deng H, Lei J, Yi F, Li J, Fan X De, et al. Efficacy of bevacizumab combined with erlotinib for advanced hepatocellular carcinoma: A single-arm meta-analysis based on prospective studies. BMC Cancer. 2019;19(1):1–13.

55. Abou-Alfa GK, Johnson P, Knox JJ, Davidenko I, Lacava J, Leung T, et al. Doxorubicin Plus Sorafenib vs Doxorubicin Alone in Patients With Advanced Hepatocellular Carcinoma. Jama. 2010;304(19):2154–60.

56. Sangro B, Gomez-Martin C, De La Mata M, Iñarrairaegui M, Garralda E, Barrera P, et al. A clinical trial of CTLA-4 blockade with tremelimumab in patients with hepatocellular carcinoma and chronic hepatitis C. J Hepatol. 2013;59(1):81–8.

57. Yau T, Hsu C, Kim TY, Choo SP, Kang YK, Hou MM, et al. Nivolumab in advanced hepatocellular carcinoma: Sorafenib-experienced Asian cohort analysis. J Hepatol. 2019;71(3):543–52.

58. Zhang L, Zhang H, Zhao Y, Li Z, Chen S, Zhai J, et al. Inhibitor selectivity between aldo-keto reductase superfamily members AKR1B10 and AKR1B1: Role of Trp112 (Trp111). FEBS Lett. 2013;587(22):3681–6.

59. Barski OA, Tipparaju SM, Bhatnagar A. The aldo-keto reductase superfamily and its role in drug metabolism and detoxification. Drug Metab Rev. 2008;40(4):553–624.

60. Mindnich RD, Penning TM. Aldo-keto reductase (AKR) superfamily: genomics and annotation. Hum Genomics. 2009;3(4):362–70.

61. Cao D, Fan ST, Chung SSM. Identification and characterization of a novel human aldose reductase- like gene. J Biol Chem. 1998;273(19):11429–35.

62. Hyndman DJ, Flynn TG. Sequence and expression levels in human tissues of a new member of the aldo-keto reductase family. Biochim Biophys Acta - Gene Struct Expr. 1998;1399(2–3):198–202.

63. Krause N, Wegner A. Fructose Metabolism in Cancer. Cells. 2020;9(12).

64. Srivastava SK, Ramana K V., Bhatnagar A. Role of aldose reductase and oxidative damage in diabetes and the consequent potential for therapeutic options. Endocr Rev. 2005;26(3):380–92.

65. Weber S, Salabei JK, Möller G, Kremmer E, Bhatnagar A, Adamski J, et al. Aldo-keto Reductase 1B15 (AKR1B15): A mitochondrial human aldo-keto reductase with activity toward steroids and 3-keto-acyl-CoA conjugates. J Biol Chem. 2015;290(10):6531–45.

66. Zeng C-M, Chang L-L, Ying M-D, Cao J, He Q-J, Zhu H, et al. Aldo–Keto Reductase AKR1C1–AKR1C4: Functions, Regulation, and Intervention for Anti-cancer Therapy. Front Pharmacol. 2017;8(119):1–9.

67. Jez JM, Bennett MJ, Schlegel BP, Lewis M, Penning TM. Comparative anatomy of the aldo-keto reductase superfamily. Biochem J. 1997;326(3):625–36.

68. Jin Y, Duan L, Lee SH, Kloosterboer HJ, Blair IA, Penning TM. Human cytosolic hydroxysteroid dehydrogenases of the aldo-ketoreductase superfamily catalyze reduction of conjugated steroids: Implications for phase I and phase II steroid hormone metabolism. J Biol Chem. 2009;284(15):10013–22.

69. Rižner TL, Penning TM. Role of aldo-keto reductase family 1 (AKR1) enzymes in human steroid metabolism. Steroids. 2014;79:49–63.

70. Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, et al. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Sci. 2021;30(1):70–82.

71. Penning TM. The aldo-keto reductases (AKRs): Overview. Chem Biol Interact. 2015;234:236–46.

72. Shen Y, Zhong L, Johnson S, Cao D. Human aldo-keto reductases 1B1 and 1B10: A comparative study on their enzyme activity toward electrophilic carbonyl compounds. Chem Biol Interact. 2011;191(1–3):192–8.

73. Zhong L, Liu Z, Yan R, Johnson S, Zhao Y, Fang X, et al. Aldo-keto reductase family 1 B10 protein detoxifies dietary and lipid-derived alpha, beta-unsaturated carbonyls at physiological levels. Biochem Biophys Res Commun. 2009;387(2):245–50.

74. Cooper WC, Jin Y, Penning TM. Elucidation of a complete kinetic mechanism for a mammalian hydroxysteroid dehydrogenase (HSD) and identification of all enzyme forms on the reaction coordinate: The example of rat liver 3α-HSD (AKR1C9). J Biol Chem. 2007;282(46):33484–93.

75. Higaki Y, Usami N, Shintani S, Ishikura S, El-Kabbani O, Hara A. Selective and potent inhibitors of human 20α-hydroxysteroid dehydrogenase (AKR1C1) that metabolizes neurosteroids derived from progesterone. Chem Biol Interact. 2003;143–144:503–13.

76. Jeon SM, Chandel NS, Hay N. AMPK regulates NADPH homeostasis to promote tumour cell survival during energy stress. Nature. 2012;485(7400):661–5.

77. Ruiz FX, Porté S, Parés X, Farrés J. Biological role of aldo-keto reductases in retinoic acid biosynthesis and signaling. Front Pharmacol. 2012;3 APR(April):1–13.

78. Endo S, Matsunaga T, Mamiya H, Ohta C, Soda M, Kitade Y, et al. Kinetic studies of AKR1B10, human aldose reductase-like protein: Endogenous substrates and inhibition by steroids. Arch Bioch Biophys. 2009;487(1):1-9.

79. Matsunaga T, Wada Y, Endo S, Soda M, El-Kabbani O, Hara A. Aldo-keto reductase 1B10 and its role in proliferation capacity of drug-resistant cancers. Front Pharmacol. 2012;3 JAN(January):1–11.

80. Ma J, Yan R, Zu X, Cheng JM, Rao K, Liao DF, et al. Aldo-keto reductase family 1 B10 affects fatty acid synthesis by regulating the stability of acetyl-CoA carboxylase-α in breast cancer cells. J Biol Chem. 2008;283(6):3418–23.

81. Huang L, He R, Luo W, Zhu Y-S, Li J, Tan T, et al. Aldo-Keto Reductase Family 1 Member B10 Inhibitors: Potential Drugs for Cancer Treatment. Recent Pat Anticancer Drug Discov. 2016;11(2):184–96.

82. Taskoparan B, Seza EG, Demirkol S, Tuncer S, Stefek M, Gure AO, et al. Opposing roles of the aldo-keto reductases AKR1B1 and AKR1B10 in colorectal cancer. Cell Oncol. 2017;40(6):563–78.

83. Canlı SD, Seza EG, Sheraj I, Gömçeli I, Turhan N, Carberry S, et al. Evaluation of an aldo-keto reductase gene signature with prognostic significance in colon cancer via activation of epithelial to mesenchymal transition and the p70S6K pathway. 2020;41(9):1219–1228.

84. Ohashi T, Idogawa M, Sasaki Y, Suzuki H, Tokino T. AKR1B10, a transcriptional target of p53, is downregulated in colorectal cancers associated with poor prognosis. Mol Cancer Res. 2013;11(12):1554–63.

85. Yan R, Zu X, Ma J, Liu Z, Adeyanju M, Cao D. Aldo-keto reductase family 1 B10 gene silencing results in growth inhibition of colorectal cancer cells: Implication for cancer intervention. Int J Cancer. 2007;121(10):2301–6.

86. Zu X, Yan R, Pan J, Zhong L, Cao Y, Ma J, et al. Aldo-keto reductase 1B10 protects human colon cells from DNA damage induced by electrophilic carbonyl compounds. Mol Carcinog. 2017;56(1):118–29.

87. Ha SY, Song DH, Lee JJ, Lee HW, Cho SY, Park CK. High expression of aldo-keto reductase 1B10 is an independent predictor of favorable prognosis in patients with hepatocellular carcinoma. Gut Liver. 2014;8(6):648–54.

88. Jin J, Liao W, Yao W, Zhu R, Li Y, He S. Aldo-keto Reductase Family 1 Member B 10 Mediates Liver Cancer Cell Proliferation through Sphingosine-1-Phosphate. Sci Rep. 2016;6(October 2015):1–11.

89. Ye X, Li C, Zu X, Lin M, Liu Q, Liu J, et al. A Large-Scale Multicenter Study Validates Aldo-Keto Reductase Family 1 Member B10 as a Prevalent Serum Marker for Detection of Hepatocellular Carcinoma. Hepatology. 2019;69(6):2489–501.

90. Cheng BY, Lau EY, Leung HW, Leung CON, Ho NP, Gurung S, et al. Irak1 augments cancer stemness and drug resistance via the ap-1/akr1b10 signaling cascade in hepatocellular carcinoma. Cancer Res. 2018;78(9):2332–42.

91. Distefano JK, Davis B. Diagnostic and prognostic potential of akr1b10 in human hepatocellular carcinoma. Cancers (Basel). 2019;11(4):1–13.

92. Reddy KA, Kumar PU, Srinivasulu M, Triveni B, Sharada K, Ismail A, et al. Overexpression and enhanced specific activity of aldoketo reductases (AKR1B1 & AKR1B10) in human breast cancers. The Breast. 2017;31:137–43.

93. van Weverwijk A, Koundouros N, Iravani M, Ashenden M, Gao Q, Poulogiannis G, et al. Metabolic adaptability in metastatic breast cancer by AKR1B10-dependent balancing of glycolysis and fatty acid oxidation. Nat Commun. 2019;10(1).

94. Hung JJ, Yeh YC, Hsu WH. Prognostic significance of AKR1B10 in patients with resected lung adenocarcinoma. Thorac Cancer. 2018;9(11):1492–9.

95. Cho, E., Cha, Y., Kim, HS., Kim, NH., Yook J. The pentose phosphate pathway as a potential target for cancer therapy. Vol. 26, Biomolecules and Therapeutics. Korean Society of Applied Pharmacology; 2018. p. 29–38.

96. Waitkus MS, Diplas BH, Yan H. Isocitrate dehydrogenase mutations in gliomas. Neuro Oncol. 2016;18(1):16–26.

97. Kong MJ, Han SJ, Kim JI, Park JW, Park KM. Mitochondrial NADP+-dependent isocitrate dehydrogenase deficiency increases cisplatin-induced oxidative damage in the kidney tubule cells article. Cell Death Dis. 2018;9(5).

98. Jiang Y, Sun A, Zhao Y, Ying W, Sun H, Yang X, et al. Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. Nature. 2019;567(7747):257–61.

99. Tomalik-Scharte D, Lazar A, Fuhr U, Kirchheiner J. The clinical role of genetic polymorphisms in drug-metabolizing enzymes. Pharmacogenomics J. 2008;8(1):4–15.

100. Terunuma A, Putluri N, Mishra P, Mathé EA, Dorsey TH, Yi M, et al. MYC-driven accumulation of 2-hydroxyglutarate is associated with breast cancer prognosis. J Clin Invest. 2014;124(1):398–412.

101. Peng X, Chen Z, Farshidfar F, Xu X, Lorenzi PL, Wang Y, et al. Molecular Characterization and Clinical Relevance of Metabolic Expression Subtypes in Human Cancers. Cell Rep. 2018;23(1):255-269.e4.

102. Rosario SR, Long MD, Affronti HC, Rowsam AM, Eng KH, Smiraglia DJ. Pan-cancer analysis of transcriptional metabolic dysregulation using The Cancer Genome Atlas. Nat Commun. 2018 Dec 1;9(1).

103. Sun C, Li T, Song X, Huang L, Zang Q, Xu J, et al. Spatially resolved metabolomics to discover tumor-associated metabolic alterations. Proc Natl Acad Sci U S A. 2019;116(1):52–7.

104. Hu J, Locasale JW, Bielas JH, O'Sullivan J, Sheahan K, Cantley LC, et al. Heterogeneity of tumor-induced gene expression changes in the human metabolic network. Nat Biotechnol. 2013;31(6):522–9.

105. Wurmbach E, Chen YB, Khitrov G, Zhang W, Roayaie S, Schwartz M, et al. Genome-wide molecular profiles of HCV-induced dysplasia and hepatocellular carcinoma. Hepatology. 2007;45(4):938–47.

106. Grinchuk O V., Yenamandra SP, Iyer R, Singh M, Lee HK, Lim KH, et al. Tumor-adjacent tissue co-expression profile analysis reveals pro-oncogenic ribosomal gene signature for prognosis of resectable hepatocellular carcinoma. Mol Oncol. 2018;12(1):89–113.

107. Lee JS, Heo J, Libbrecht L, Chu IS, Kaposi-Novak P, Calvisi DF, et al. A novel prognostic subtype of human hepatocellular carcinoma derived from hepatic progenitor cells. Nat Med. 2006;12(4):410–6.

108. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016;44(8):e71.

109. Mounir M, Lucchetta M, Silva TC, Olsen C, Bontempi G, Chen X, et al. New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEX. PLoS Comput Biol. 2019;15(3):e1006701.

110. Anders S, Pyl PT, Huber W. HTSeq-A Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166–9.

111. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):1–21.

112. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;1–12.

113. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30(10):1363–9.

114. Touleimat N, Tost J. Complete pipeline for Infinium® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. Epigenomics. 2012;4(3):325–41.

115. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010;11.

116. Gautier L, Cope L, Bolstad BM, Irizarry RA. Affy - Analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20(3):307–15.

117. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4(2):249–264.

118. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.

119. Boellner S, Becker K-F. Reverse Phase Protein Arrays—Quantitative Assessment of Multiple Biomarkers in Biopsies for Clinical Use. Microarrays. 2015;4(2):98–114.

120. Yu G, Wang LG, Han Y, He QY. ClusterProfiler: An R package for comparing biological themes among gene clusters. Omi A J Integr Biol. 2012;16(5):284–7.

121. Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, Karp PD. Computational prediction of human metabolic pathways from the complete human genome. Genome Biol. 2005;6(1):1–17.

122. Murtagh F, Legendre P. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? J Classif. 2014;3:274–95.

123. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. J Stat Softw. 2011;39(5):1–13.

124. Mueckler M, Thorens B. The SLC2 (GLUT) family of membrane transporters. Mol Aspects Med. 2013;34(2–3):121–38.

125. Da Veiga Moreira J, Hamraz M, Abolhassani M, Bigan E, Pérès S, Paulevé L, et al. The redox status of cancer cells supports mechanisms behind the Warburg effect. Metabolites. 2016;6(4):1–12.

126. Tanner LB, Goglia AG, Wei MH, White E, Toettcher JE, Rabinowitz JD, et al. Four Key Steps Control Glycolytic Flux in Mammalian Cells. Cell Syst. 2018;7(1):49-62.e8.

127. Lewis CA, Parker SJ, Fiske BP, McCloskey D, Gui DY, Green CR, et al. Tracing Compartmentalized NADPH Metabolism in the Cytosol and Mitochondria of Mammalian Cells. Mol Cell. 2014;55(2):253–63.

128. Akella NM, Ciraku L, Reginato MJ. Fueling the fire: Emerging role of the hexosamine biosynthetic pathway in cancer. BMC Biol. 2019;17(1):1–14.

129. De Berardinis RJ, Chandel NS. Fundamentals of cancer metabolism. Sci Adv. 2016;2(5):(e1600200).

130. Currie E, Schulze A, Zechner R, Walther TC, Farese R V. Cellular fatty acid metabolism and cancer. Cell Metab. 2013;18(2):153–61.

131. Carracedo A, Cantley LC, Pandolfi PP. Cancer metabolism: Fatty acid oxidation in the limelight. Nat Rev Cancer. 2013;13(4):227–32.

132. Weber DD, Aminzadeh-Gohari S, Tulipan J, Catalano L, Feichtinger RG, Kofler B. Ketogenic diet in the treatment of cancer – Where do we stand? Mol Metab. 2020;33(July 2019):102–21.

133. Fafournoux P, Bruhat A, Jousse C. Amino acid regulation of gene expression. Biochem J. 2000;14(351):1–12.

134. Van Sluijters DA, Dubbelhuis PF, Blommaart EFC, Meijer AJ. Amino-acid-dependent signal transduction. Biochem J. 2000;351(3):545–50.

135. Mayers JR, Vander Heiden MG. Famine versus feast: Understanding the metabolism of tumors in vivo. Trends Biochem Sci. 2015;40(3):130–40.

136. Vettore L, Westbrook RL, Tennant DA. New aspects of amino acid metabolism in cancer. Br J Cancer. 2020;122(2):150–6.

137. Keshet R, Szlosarek P, Carracedo A, Erez A. Rewiring urea cycle metabolism in cancer to support anabolism. Nat Rev Cancer. 2018;18(10):634–45.

138. Eniafe J, Jiang S. The functional roles of TCA cycle metabolites in cancer. Oncogene. 2021;40(19):3351–63.

139. Vacanti NM, Divakaruni AS, Green CR, Parker SJ, Henry RR, Ciaraldi TP, et al. Regulation of substrate utilization by the mitochondrial pyruvate carrier. Mol Cell. 2014;56(3):425–35.

140. Baffy G. Mitochondrial uncoupling in cancer cells: Liabilities and opportunities. Biochim Biophys Acta - Bioenerg. 2017;1858(8):655–64.

141. Sreedhar A, Cassell T, Smith P, Lu D, Nam HW, Lane AN, et al. UCP2 overexpression redirects glucose into anabolic metabolic pathways. Proteomics. 2019;19:e1800353.

142. Zhao RZ, Jiang S, Zhang L, Yu Z Bin. Mitochondrial electron transport chain, ROS generation and uncoupling (Review). Int J Mol Med. 2019;44(1):3–15.

143. Levental KR, Malmberg E, Symons JL, Fan YY, Chapkin RS, Ernst R, et al. Lipidomic and biophysical homeostasis of mammalian membranes counteracts dietary lipid perturbations to maintain cellular fitness. Nat Commun. 2020;11(1):1–13.

144. Rui L. Energy metabolism in the liver. Compr Physiol. 2014;4(1):177–97.

145. Trefts E, Gannon M, Wasserman DH. The liver. Curr Biol. 2017;27(21):R1147–51.

146. Nguyen J, Le QH, Duong BH, Sun P, Pham HT, Ta VT, et al. A Matched Case-Control Study of Risk Factors for Breast Cancer Risk in Vietnam. Int J Breast Cancer. 2016;1–7.

147. Buffa FM, Harris AL, West CM, Miller CJ. Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene. Br J Cancer. 2010;102(2):428–35.

148. Ragnum HB, Vlatkovic L, Lie AK, Axcrona K, Julin CH, Frikstad KM, et al. The tumour hypoxia marker pimonidazole reflects a transcriptional programme associated with aggressive prostate cancer. Br J Cancer. 2015;112(2):382–90.

149. Winter SC, Buffa FM, Silva P, Miller C, Valentine HR, Turley H, et al. Relation of a hypoxia metagene derived from head and neck cancer to prognosis of multiple cancers. Cancer Res. 2007;67(7):3441–9.

150. Fujita M, Yamaguchi R, Hasegawa T, Shimada S, Arihiro K, Hayashi S, et al. Classification of primary liver cancer with immunosuppression mechanisms and correlation with genomic alterations. EBioMedicine. 2020;53:102659.

151. Prevarskaya N, Skryma R, Shuba Y. Ion channels in cancer: Are cancer hallmarks oncochannelopathies? Physiol Rev. 2018;98(2):559–621.

152. Guengerich FP, Waterman MR, Egli M. Recent Structural Insights into Cytochrome P450 Function. Trends Pharmacol Sci. 2016;37(8):625–40.

153. Manikandan P, Nagini S. Cytochrome P450 Structure, Function and Clinical Significance: A Review. Curr Drug Targets. 2017;19(1):38–54.

154. Pappa A, Estey T, Manzer R, Brown D, Vasiliou V. Human aldehyde dehydrogenase 3A1 (ALDH3A1): Biochemical characterization and immunohistochemical localization in the cornea. Biochem J. 2003;376(3):615–23.

155. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2(5):401–4.

156. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013;6(269).

157. Tan H, Bao J, Zhou X. Genome-wide mutational spectra analysis reveals significant cancer-specific heterogeneity. Sci Rep. 2015;5:1–14.

158. Mantovani F, Collavin L, Del Sal G. Mutant p53 as a guardian of the cancer cell. Cell Death Differ. 2019;26(2):199–212.

159. Kim S, Jeong S. Mutation hotspots in the β-catenin gene: Lessons from the human cancer genome databases. Mol Cells. 2019;42(1):8–16.

160. Preissner SC, Hoffmann MF, Preissner R, Dunkel M, Gewiess A, Preissner S. Polymorphic cytochrome P450 enzymes (CYPs) and their role in personalized therapy. PLoS One. 2013;8(12):1–12.

161. Ferrari P, McKay JD, Jenab M, Brennan P, Canzian F, Vogel U, et al. Alcohol dehydrogenase and aldehyde dehydrogenase gene polymorphisms, alcohol intake and the risk of colorectal cancer in the European Prospective Investigation into Cancer and Nutrition study. Eur J Clin Nutr. 2012;66(12):1303–8.

162. Xu H, Yan M, Patra J, Natrajan R, Yan Y, Swagemakers S, et al. Enhanced RAD21 cohesin expression confers poor prognosis and resistance to chemotherapy in high grade luminal, basal and HER2 breast cancers. Breast Cancer Res. 2011;13(1):1–10.

163. Ropero S, Setien F, Espada J, Fraga MF, Herranz M, Asp J, et al. Epigenetic loss of the familial tumor-suppressor gene exostosin-1 (EXT1) disrupts heparan sulfate synthesis in cancer cells. Hum Mol Genet. 2004;13(22):2753–65.

164. Xu F, Gu J, Wang L, Liu R, Yuan Y, Wang H, et al. Up-regulation of EIF3e is associated with the progression of esophageal squamous cell carcinoma and poor prognosis in patients. J Cancer. 2018;9(7):1135–44.

165. Kulis M, Esteller M. DNA Methylation and Cancer. Adv Genet. 2010;70(C):27–56.

166. Jones PA. Functions of DNA methylation: Islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012;13(7):484–92.

167. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS. 2005;102(43):15545–50.

168. Maryanovich M, Gross A. A ROS rheostat for cell fate regulation. Trends Cell Biol. 2013;23(3):129–34.

169. Perillo B, Di Donato M, Pezone A, Di Zazzo E, Giovannelli P, Galasso G, et al. ROS in cancer therapy: the bright side of the moon. Exp Mol Med. 2020;52(2):192–203.

170. Guillemette C, Lévesque É, Rouleau M. Pharmacogenomics of human uridine diphospho-glucuronosyltransferases and clinical implications. Clin Pharmacol Ther. 2014;96(3):324–39.

171. Allain EP, Rouleau M, Lévesque E, Guillemette C. Emerging roles for UDP-glucuronosyltransferases in drug resistance and cancer progression. Br J Cancer. 2020;122(9):1277–87.

172. Lévesque E, Labriet A, Hovington H, Allain ÉP, Melo-Garcia L, Rouleau M, et al. Alternative promoters control UGT2B17-dependent androgen catabolism in prostate cancer and its influence on progression. Br J Cancer. 2020;122(7):1068–76.

173. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Syst. 2015;1(6):417–25.

174. Zou Z, Tao T, Li H, Zhu X. MTOR signaling pathway and mTOR inhibitors in cancer: Progress and challenges. Cell Biosci. 2020;10(1):1–11.

175. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep learning–based multi-omics integration robustly predicts survival in liver cancer. Clin Cancer Res. 2018;24(6):1248–59.

176. Fahrmann JF, Grapov D, Wanichthanarak K, DeFelice BC, Salemi MR, Rom WN, et al. Integrated metabolomics and proteomics highlight altered nicotinamide and polyamine pathways in lung adenocarcinoma. Carcinogenesis. 2017;38(3):271–80.

177. Becker D, Sfakianakis I, Krupp M, Staib F, Gerhold-Ay A, Victor A, et al. Genetic signatures shared in embryonic liver development and liver cancer define prognostically relevant subgroups in HCC. Mol Cancer. 2012;11:1–11.

178. Cassandri M, Smirnov A, Novelli F, Pitolli C, Agostini M, Malewicz M, et al. Zinc-finger proteins in health and disease. Cell Death Discov. 2017;3(1).

179. Ichimura Y, Komatsu M. Activation of p62/SQSTM1-keap1-nuclear factor erythroid 2-related factor 2 pathway in cancer. Front Oncol. 2018;8(JUN):1–8.

180. Saito T, Ichimura Y, Taguchi K, Suzuki T, Mizushima T, Takagi K, et al. P62/Sqstm1 promotes malignancy of HCV-positive hepatocellular carcinoma through Nrf2-dependent metabolic reprogramming. Nat Commun. 2016;7(May):1–16.

181. Chen R, Zhao WQ, Fang C, Yang X, Ji M. Histone methyltransferase SETD2: A potential tumor suppressor in solid cancers. J Cancer. 2020;11(11):3349–56.

182. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A Software Environment for Integrated Models. Genome Res. 2003;13(22):2498–504.

183. Rao C V., Asch AS, Yamada HY. Frequently mutated genes/pathways and genomic instability as prevention targets in liver cancer. Carcinogenesis. 2017;38(1):2–11.

184. Masoomian B, Shields JA, Shields CL. Overview of BAP1 cancer predisposition syndrome and the relationship to uveal melanoma. J Curr Ophthalmol. 2018;30(2):102–9.

185. Shriver M, Stroka KM, Vitolo MI, Martin S, Huso DL, Konstantopoulos K, et al. Loss of giant obscurins from breast epithelium promotes epithelial-to-mesenchymal transition, tumorigenicity and metastasis. Oncogene. 2015;34(32):4248–59.

186. Perry NA, Ackermann MA, Shriver M, Hu LYR, Kontrogianni-Konstantopoulos A. Obscurins: Unassuming giants enter the spotlight. IUBMB Life. 2013;65(6):479–86.

187. Singh A, Boldin-Adamsky S, Thimmulappa RK, Rath SK, Ashush H, Coulter J, et al. RNAi-mediated silencing of nuclear factor erythroid-2-related factor 2 gene expression in non-small cell lung cancer inhibits tumor growth and increases efficacy of chemotherapy. Cancer Res. 2008;68(19):7975–84.

188. Li QK, Singh A, Biswal S, Askin F, Gabrielson E. KEAP1 gene mutations and NRF2 activation are common in pulmonary papillary adenocarcinoma. J Hum Genet. 2011;56(3):230–4.

189. Chang YS, Huang H Da, Yeh KT, Chang JG. Identification of novel mutations in endometrial cancer patients by whole-exome sequencing. Int J Oncol. 2017;50(5):1778–84.

190. Frank R, Scheffler M, Merkelbach-Bruse S, Ihle MA, Kron A, Rauer M, et al. Clinical and pathological characteristics of KEAP1- and NFE2L2-mutated Non–Small Cell Lung Carcinoma (NSCLC). Clin Cancer Res. 2018;24(13):3087–96.

191. Fu X, Tian M, Gu J, Cheng T, Ma D, Feng L, et al. SF3B1 mutation is a poor prognostic indicator in luminal B and progesterone receptor-negative breast cancer patients. Oncotarget. 2017;8(70):115018–27.

192. Basu A, Bodycombe NE, Cheah JH, Price E V., Liu K, Schaefer GI, et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. Cell. 2013;154(5):1151–61.

193. Seashore-Ludlow B, Rees MG, Cheah JH, Coko M, Price E V., Coletti ME, et al. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. Cancer Discov. 2015;5(11):1210–23.

194. Rees MG, Seashore-Ludlow B, Cheah JH, Adams DJ, Price E V., Gill S, et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. Nat Chem Biol. 2016;12(2):109–16.

195. Heske CM. Beyond Energy Metabolism: Exploiting the Additional Roles of NAMPT for Cancer Therapy. Front Oncol. 2020;9(January).

196. Fan QW, Cheng CK, Nicolaides TP, Hackett CS, Knight ZA, Shokat KM, et al. A dual phosphoinositide-3-kinase α/mTOR inhibitor cooperates with blockade of epidermal growth factor receptor in PTEN-mutant glioma. Cancer Res. 2007;67(17):7960–5.

197. Fan QW, Knight ZA, Goldenberg DD, Yu W, Mostov KE, Stokoe D, et al. A dual PI3 kinase/mTOR inhibitor reveals emergent efficacy in glioma. Cancer Cell. 2006;9(5):341–9.

198. Park S, Chapuis N, Bardet V, Tamburini J, Gallay N, Willems L, et al. PI-103, a dual inhibitor of Class IA phosphatidylinositide 3-kinase and mTOR, has antileukemic activity in AML. Leukemia. 2008;22(9):1698–706.

199. Zuo S, Zhang X, Wang L. A RNA sequencing-based six-gene signature for survival prediction in patients with glioblastoma. Sci Rep. 2019;9(1):1–10.

200. Chen EG, Wang P, Lou H, Wang Y, Yan H, Bi L, et al. A robust gene expression-based prognostic risk score predicts overall survival of lung adenocarcinoma patients. Oncotarget. 2018;9(6):6862–71.

201. Zhao Y, Yang SM, Jin YL, Xiong GW, Wang P, Snijders AM, et al. A Robust Gene Expression Prognostic Signature for Overall Survival in High-Grade Serous Ovarian Cancer. J Oncol. 2019;2019.

202. Yan X, Fu X, Guo ZX, Liu XP, Liu TZ, Li S. Construction and validation of an eight-gene signature with great prognostic value in bladder cancer. J Cancer. 2020;11(7):1768–79.

203. Zhou T, Cai Z, Ma N, Xie W, Gao C, Huang M, et al. A Novel Ten-Gene Signature Predicting Prognosis in Hepatocellular Carcinoma. Front Cell Dev Biol. 2020;8(July):1–11.

204. Ouyang G, Yi B, Pan G, Chen X. A robust twelve-gene signature for prognosis prediction of hepatocellular carcinoma. Cancer Cell Int. 2020;20(1):1–18.

205. Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A deep learning mammography-based model for improved breast cancer risk prediction. Radiology. 2019;292(1):60–6.

206. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. npj Digit Med. 2018;1(1):1–10.

207. Quiñonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND. Dataset Shift in Machine Learning. 1st ed. MIT Press; 2008. 248 p.

208. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010;33(1):1–22.

209. Tibshirani R, Bien J, Friedman J, Hastie T, Simon N, Taylor J, et al. Strong rules for discarding predictors in lasso-type problems. J R Stat Soc Ser B Stat Methodol. 2012;74(2):245–66.

210. Ladanie A, Schmitt AM, Speich B, Naudet F, Agarwal A, Pereira T V., et al. Clinical Trial Evidence Supporting US Food and Drug Administration Approval of Novel Cancer Therapies Between 2000 and 2016. JAMA Netw open. 2020;3(11):e2024406.

211. Hiom SC. Diagnosing cancer earlier: Reviewing the evidence for improving cancer survival. Br J Cancer. 2015;112:S1–5.

212. Harold F. The Vital Force: A Study of Bioenergetics. First. W H Freeman & Co; 1986.

213. West G. Scale: The Universal Laws of Life and Death in Organisms, Cities and Companies. 1st ed. Penguin Press; 2017. 496 p.

214. Faubert B, Li KY, Cai L, Hensley CT, Kim J, Zacharias LG, et al. Lactate Metabolism in Human Lung Tumors. Cell. 2017;171(2):358-371.e9.

215. Hui S, Ghergurovich JM, Morscher RJ, Jang C, Teng X, Lu W, et al. Glucose feeds the TCA cycle via circulating lactate. Nature. 2017;551(7678):115–8.

216. Sellers K, Fox MP, Ii MB, Slone SP, Higashi RM, Miller DM, et al. Pyruvate carboxylase is critical for non-small-cell lung cancer proliferation. J Clin Invest. 2015;125(2):687–98.

217. Mullen AR, Wheaton WW, Jin ES, Chen PH, Sullivan LB, Cheng T, et al. Reductive carboxylation supports growth in tumour cells with defective mitochondria. Nature. 2012;481(7381):385–8.

218. Costello LC, Franklin RB. "Why do tumour cells glycolyse?": From glycolysis through citrate to lipogenesis. Mol Cell Biochem. 2005;280(1–2):1–8.

219. Kanarek N, Petrova B, Sabatini DM. Dietary modifications for enhanced cancer therapy. Nature [Internet]. 2020;579(7800):507–17.

APPENDICES

A. GSEA GO Term Enrichment in AKR1B10^{High} Patients

NAME	*SIZE*	*p-value*	*FDR*
GO NUCLEOBASE CONTAINING SMALL MOLECULE INTERCONVERSION	21	0.00	0.029
GO NADP METABOLIC PROCESS	27	0.00	0.094
GO OXIDOREDUCTASE ACTIVITY ACTING ON A SULFUR GROUP OF DONORS	43	0.00	0.082
GO DISULFIDE OXIDOREDUCTASE ACTIVITY	27	0.00	0.104
GO RESPONSE TO OXYGEN RADICAL	18	0.00	0.091
GO LIPOPOLYSACCHARIDE MEDIATED SIGNALING PATHWAY	31	0.00	0.099
GO ANTIGEN PROCESSING AND PRESENTATION OF PEPTIDE ANTIGEN VIA MHC CLASS I	90	0.00	0.102
GO REGULATION OF CELLULAR AMINO ACID METABOLIC PROCESS	64	0.01	0.100
GO CELLULAR GLUCURONIDATION	22	0.00	0.100
GO POSITIVE REGULATION OF TISSUE REMODELING	22	0.00	0.123
GO PROTEASOME ACCESSORY COMPLEX	23	0.00	0.114
GO ANTIGEN PROCESSING AND PRESENTATION OF EXOGENOUS PEPTIDE ANTIGEN VIA MHC CLASS I	65	0.00	0.118
GO NIK NF KAPPAB SIGNALING	81	0.01	0.113
GO ALDO KETO REDUCTASE NADP ACTIVITY	26	0.00	0.114
GO GLUCOSE 6 PHOSPHATE METABOLIC PROCESS	21	0.00	0.119
GO OXIDOREDUCTASE ACTIVITY ACTING ON THE CH CH GROUP OF DONORS NAD OR NADP AS ACCEPTOR	24	0.00	0.118
GO REGULATION OF VIRAL ENTRY INTO HOST CELL	25	0.00	0.111
GO GLYCOSIDE METABOLIC PROCESS	16	0.00	0.111
GO PROTEASOME BINDING	16	0.00	0.106
GO PROSTANOID METABOLIC PROCESS	26	0.00	0.102
GO GLUCURONOSYLTRANSFERASE ACTIVITY	31	0.00	0.109
GO TUMOR NECROSIS FACTOR MEDIATED SIGNALING PATHWAY	115	0.01	0.124
GO DETOXIFICATION	69	0.00	0.125
GO REGULATION OF CELLULAR AMINE METABOLIC PROCESS	84	0.01	0.134
GO TOLL LIKE RECEPTOR SIGNALING PATHWAY	84	0.00	0.129
GO ACTIVATION OF INNATE IMMUNE RESPONSE	199	0.00	0.129
GO REGULATION OF B CELL APOPTOTIC PROCESS	18	0.01	0.150
GO RESPONSE TO INTERFERON BETA	20	0.00	0.159
GO PROTEIN DISULFIDE OXIDOREDUCTASE ACTIVITY	20	0.00	0.163
GO ORGANELLAR LARGE RIBOSOMAL SUBUNIT	29	0.00	0.180

Table (Cont'd)

GO PROTEASOME COMPLEX	74	0.03	0.177
GO FLAVONOID METABOLIC PROCESS	28	0.00	0.174
GO QUINONE METABOLIC PROCESS	28	0.00	0.180
GO OXIDOREDUCTASE ACTIVITY ACTING ON THE CH OH GROUP OF DONORS NAD OR NADP AS ACCEPTOR	110	0.00	0.182
GO INNATE IMMUNE RESPONSE ACTIVATING CELL SURFACE RECEPTOR SIGNALING PATHWAY	104	0.02	0.199
GO URONIC ACID METABOLIC PROCESS	27	0.00	0.198
GO VACUOLAR ACIDIFICATION	15	0.01	0.209
GO REGULATION OF INTERLEUKIN 1 SECRETION	31	0.01	0.204
GO TOLL LIKE RECEPTOR 4 SIGNALING PATHWAY	18	0.01	0.211
GO RESPONSE TO TYPE I INTERFERON	53	0.03	0.207
GO OMEGA PEPTIDASE ACTIVITY	17	0.00	0.205
GO REGULATION OF INTERLEUKIN 8 SECRETION	17	0.01	0.213
GO NEGATIVE REGULATION OF VIRAL ENTRY INTO HOST CELL	17	0.01	0.210
GO POSITIVE REGULATION OF INNATE IMMUNE RESPONSE	237	0.01	0.223
GO GLYCERALDEHYDE 3 PHOSPHATE METABOLIC PROCESS	16	0.01	0.231
GO ALCOHOL DEHYDROGENASE NADP ACTIVITY	16	0.01	0.236
GO PH REDUCTION	37	0.00	0.236
GO SUPEROXIDE METABOLIC PROCESS	29	0.00	0.234
GO POSITIVE REGULATION OF INTERLEUKIN 1 SECRETION	24	0.01	0.232
GO REGULATION OF PROTEIN UBIQUITINATION INVOLVED IN UBIQUITIN DEPENDENT PROTEIN CATABOLIC PROCESS	100	0.04	0.231
GO GLYCOLIPID BINDING	19	0.01	0.233
GO NEGATIVE REGULATION OF VIRAL RELEASE FROM HOST CELL	16	0.02	0.238

B. GSEA KEGG Pathway Enrichment in AKR1B10^{High} Patients

NAME	*SIZE*	*p-value*	*FDR*
KEGG PENTOSE AND GLUCURONATE INTERCONVERSIONS	27	0.00	0.01
KEGG ASCORBATE AND ALDARATE METABOLISM	25	0.00	0.01
KEGG PORPHYRIN AND CHLOROPHYLL METABOLISM	40	0.00	0.02
KEGG PROTEASOME	42	0.00	0.04
KEGG GLUTATHIONE METABOLISM	47	0.00	0.05
KEGG AMINO SUGAR AND NUCLEOTIDE SUGAR METABOLISM	44	0.01	0.09
KEGG LYSOSOME	119	0.02	0.17
KEGG OTHER GLYCAN DEGRADATION	15	0.03	0.19
KEGG METABOLISM OF XENOBIOTICS BY CYTOCHROME P450	69	0.03	0.19
KEGG DRUG METABOLISM OTHER ENZYMES	51	0.02	0.17
KEGG STEROID HORMONE BIOSYNTHESIS	55	0.03	0.18
KEGG CYTOSOLIC DNA SENSING PATHWAY	40	0.04	0.19
KEGG AMINOACYL TRNA BIOSYNTHESIS	41	0.06	0.19
KEGG PENTOSE PHOSPHATE PATHWAY	26	0.03	0.20

C. GSEA Reactome Enrichment in AKR1B10^{High} Patients

Name	*SIZE*	*p-val*	*FDR*
REACTOME SYNTHESIS AND INTERCONVERSION OF NUCLEOTIDE DI AND TRIPHOSPHATES	18	0.00	0.00
REACTOME REGULATION OF ORNITHINE DECARBOXYLASE ODC	48	0.00	0.08
REACTOME DESTABILIZATION OF MRNA BY AUF1 HNRNP D0	51	0.00	0.06
REACTOME METABOLISM OF NUCLEOTIDES	70	0.00	0.07
REACTOME GLUCURONIDATION	18	0.01	0.08
REACTOME REGULATION OF MRNA STABILITY BY PROTEINS THAT BIND AU RICH ELEMENTS	82	0.02	0.08
REACTOME AUTODEGRADATION OF THE E3 UBIQUITIN LIGASE COP1	48	0.01	0.08
REACTOME VIF MEDIATED DEGRADATION OF APOBEC3G	48	0.01	0.08
REACTOME ER PHAGOSOME PATHWAY	60	0.01	0.07
REACTOME ANTIGEN PROCESSING CROSS PRESENTATION	74	0.01	0.07
REACTOME CROSS PRESENTATION OF SOLUBLE EXOGENOUS ANTIGENS ENDOSOMES	47	0.01	0.06
REACTOME CYTOSOLIC TRNA AMINOACYLATION	24	0.01	0.08
REACTOME GOLGI ASSOCIATED VESICLE BIOGENESIS	50	0.01	0.08
REACTOME AUTODEGRADATION OF CDH1 BY CDH1 APC C	57	0.02	0.07
REACTOME ACTIVATION OF NF KAPPAB IN B CELLS	62	0.02	0.07
REACTOME TRANS GOLGI NETWORK VESICLE BUDDING	57	0.01	0.07
REACTOME P53 INDEPENDENT G1 S DNA DAMAGE CHECKPOINT	49	0.02	0.06
REACTOME TRNA AMINOACYLATION	42	0.02	0.07
REACTOME IRON UPTAKE AND TRANSPORT	35	0.01	0.07
REACTOME PHASE II CONJUGATION	68	0.00	0.07
REACTOME REGULATION OF APOPTOSIS	57	0.03	0.07
REACTOME CDK MEDIATED PHOSPHORYLATION AND REMOVAL OF CDC6	47	0.03	0.07
REACTOME SCF BETA TRCP MEDIATED DEGRADATION OF EMI1	50	0.03	0.07
REACTOME SIGNALING BY WNT	63	0.03	0.07
REACTOME ACTIVATED TAK1 MEDIATES P38 MAPK ACTIVATION	17	0.01	0.07
REACTOME TAK1 ACTIVATES NFKB BY PHOSPHORYLATION AND ACTIVATION OF IKKS COMPLEX	23	0.01	0.11
REACTOME CDT1 ASSOCIATION WITH THE CDC6 ORC ORIGIN COMPLEX	55	0.04	0.11
REACTOME INTERFERON ALPHA BETA SIGNALING	50	0.03	0.11
REACTOME CYCLIN E ASSOCIATED EVENTS DURING G1 S TRANSITION	63	0.04	0.11

Table (Cont'd)

REACTOME P53 DEPENDENT G1 DNA DAMAGE RESPONSE	54	0.05	0.11
REACTOME APC C CDC20 MEDIATED DEGRADATION OF MITOTIC PROTEINS	66	0.04	0.11
REACTOME MEMBRANE TRAFFICKING	123	0.01	0.11
REACTOME DOWNSTREAM SIGNALING EVENTS OF B CELL RECEPTOR BCR	93	0.03	0.12
REACTOME SIGNALING BY THE B CELL RECEPTOR BCR	122	0.02	0.11
REACTOME SCFSKP2 MEDIATED DEGRADATION OF P27 P21	54	0.05	0.13
REACTOME APC C CDH1 MEDIATED DEGRADATION OF CDC20 AND OTHER APC C CDH1 TARGETED PROTEINS IN LATE MITOSIS EARLY G1	65	0.07	0.14
REACTOME NUCLEOTIDE BINDING DOMAIN LEUCINE RICH REPEAT CONTAINING RECEPTOR NLR SIGNALING PATHWAYS	44	0.02	0.14
REACTOME INSULIN RECEPTOR RECYCLING	21	0.03	0.15
REACTOME ASSEMBLY OF THE PRE REPLICATIVE COMPLEX	64	0.06	0.15
REACTOME DESTABILIZATION OF MRNA BY TRISTETRAPROLIN TTP	17	0.04	0.15
REACTOME NEF MEDIATES DOWN MODULATION OF CELL SURFACE RECEPTORS BY RECRUITING THEM TO CLATHRIN ADAPTERS	21	0.04	0.20
REACTOME ANTIGEN PRESENTATION FOLDING ASSEMBLY AND PEPTIDE LOADING OF CLASS I MHC	21	0.08	0.20
REACTOME HOST INTERACTIONS OF HIV FACTORS	120	0.07	0.23
REACTOME THE ROLE OF NEF IN HIV1 REPLICATION AND DISEASE PATHOGENESIS	28	0.06	0.23

D. LASSO Model Clinical Dependencies

Variables		*Two Groups*			*High and Low*		
Characteristic	Variables	High	Low	p-value	High	Low	p-value
Gender	Male	81	161	0.085	81	61	0.62
	Female	28	89		28	26	
Age	<= 60	50	120	0.8	50	36	0.63
	>60	59	130		59	51	
Grade	I	34	134	0.0013	34	50	0.001
	II	33	50		33	13	
	III	33	49		33	16	
	IV	1	4		1	0	
Neoplasm Grade	G1	10	44	0.0051	10	31	<0.0001
	G2	45	125		45	41	
	G3	48	70		48	13	
	G4	6	6		6	1	
TNM Stage	T1	35	142	< 0.0001	35	54	0.0002
	T2	36	51		36	13	
	T3	17	26		17	10	
	T3A	11	17		11	6	
	T4	6	7		6	0	
TP53	WT	46	190	< 0.0001	77	31	<0.0001
	MUT	62	52		55	29	
CTNNB1	WT	77	176	0.88	59	66	0.48
	MUT	31	66		28	21	
ALB	WT	93	209	1	93	71	0.92
	MUT	15	33		15	13	
KEAP1	WT	99	233	0.12	99	81	0.29
	MUT	9	9		9	3	
NFE2L2	WT	99	227	0.62	99	80	0.49
	MUT	9	15		9	4	

E. Expression Differences of Glycolysis-Related Genes

Table showing expression of genes involved in glycolysis, PPP and Hexose transport in the cell. Positive LFC shows upregulation in low-risk group and negative LFC upregulation in high-risk group.

Symbol	*Pathway*	*4-Gene Model lfc*	*2-Gene Model lfc*
ADH1A	Glycolysis	1.605	1.510
ADH1B	Glycolysis	2.512	2.023
ADH1C	Glycolysis	1.583	1.579
ADH4	Glycolysis	2.141	2.005
ADH5	Glycolysis	0.472	0.425
ADH6	Glycolysis	1.208	1.236
ADH7	Glycolysis	0.655	0.482
ADHFE1	Glycolysis	1.380	1.054
AKR1A1	Glycolysis	0.036	0.099
ALDH1A3	Glycolysis	1.796	0.955
ALDH1B1	Glycolysis	0.561	0.924
ALDH2	Glycolysis	1.407	1.520
ALDH3A1	Glycolysis	0.017	-0.532
ALDH3A2	Glycolysis	0.612	0.502
ALDH3B1	Glycolysis	-0.646	-0.996
ALDH3B2	Glycolysis	-0.409	-2.048
ALDH7A1	Glycolysis	0.899	0.896
ALDH9A1	Glycolysis	0.627	0.748
ALDOA	Glycolysis	-1.650	-1.505
ALDOB	Glycolysis	1.363	1.633
ALDOC	Glycolysis	-0.071	-0.120
BPGM	Glycolysis	0.023	0.031
DERA	Glycolysis	0.322	0.354
ENO1	Glycolysis	-1.462	-1.201
ENO2	Glycolysis	-2.078	-2.245
ENO3	Glycolysis	0.352	0.886
FBP1	Glycolysis	1.761	1.680
FBP2	Glycolysis	-0.073	0.130
G6PC	Glycolysis	2.305	2.233
G6PC2	Glycolysis	1.199	1.290
GALM	Glycolysis	0.334	0.243
GAPDH	Glycolysis	-1.002	-0.856
GPI	Glycolysis	-0.485	-0.135
HK1	Glycolysis	-0.378	0.240
HK2	Glycolysis	-0.924	-1.829
HK3	Glycolysis	-1.372	-1.184
HK4	Glycolysis	1.261	2.029

Table (Cont'd)

HKDC1	Glycolysis	-0.799	-1.391
LDHA	Glycolysis	-0.413	-0.177
LDHAL6A	Glycolysis	-0.667	-0.459
LDHAL6B	Glycolysis	0.270	0.156
LDHB	Glycolysis	0.041	-0.618
LDHC	Glycolysis	-0.047	0.313
MCT1	Glycolysis	0.005	0.269
MCT2	Glycolysis	-0.122	-0.015
MCT3	Glycolysis	0.278	0.433
MTC4	Glycolysis	-2.603	-2.575
PC	Glycolysis	0.903	0.907
PCK1	Glycolysis	2.130	1.791
PCK2	Glycolysis	1.682	1.489
PFKL	Glycolysis	0.231	0.199
PFKM	Glycolysis	-0.747	-0.199
PFKP	Glycolysis	-1.889	-1.707
PGAM1	Glycolysis	-0.249	-0.089
PGAM4	Glycolysis	0.061	0.295
PGK1	Glycolysis	-0.545	-0.417
PGK2	Glycolysis	-2.212	-1.547
PGM1	Glycolysis	0.567	0.770
PGM2	Glycolysis	0.072	0.561
PKLR	Glycolysis	0.931	1.075
PKM	Glycolysis	-2.476	-2.447
6PGL	PPP	-0.229	-0.275
G6PD	PPP	-3.874	-3.457
PGD	PPP	-1.547	-1.154
PRPS1	PPP	0.004	0.159
PRPS1L1	PPP	-2.834	-2.346
PRPS2	PPP	0.118	-0.174
RBKS	PPP	0.492	0.533
RPE	PPP	-0.283	-0.250
RPEL1	PPP	-0.411	-0.076
RPIA	PPP	-0.512	-0.536
SHPK	PPP	0.685	0.756
TALDO1	PPP	-0.840	-0.857
TKT	PPP	-1.561	-1.641
TPI1	PPP	-0.437	-0.399
SLC2A1	Transport	-2.561	-2.143
SLC2A2	Transport	1.806	1.520
SLC2A3	Transport	-0.211	-0.048
SLC2A4	Transport	0.636	1.231
SLC2A5	Transport	-1.844	-0.641
SLC2A6	Transport	-1.136	-1.074
SLC2A7	Transport	-0.679	-0.737

Table (Cont'd)

SLC2A8	Transport	0.369	0.431
SLC2A9	Transport	0.618	0.824
SLC2A10	Transport	1.037	0.789
SLC2A11	Transport	0.040	0.184
SLC2A12	Transport	0.196	0.472
SLC2A13	Transport	-0.184	-0.171
SLC2A14	Transport	0.627	0.693
SLC45A1	Transport	0.880	0.731
SLC45A2	Transport	-0.261	0.324
SLC45A3	Transport	0.623	0.319
SLC45A4	Transport	0.645	-0.445
SLC50A1	Transport	-0.289	-0.530
SLC5A1	Transport	1.818	-0.215
SLC5A10	Transport	-0.258	0.105
SLC5A11	Transport	-2.049	-2.508
SLC5A2	Transport	-0.228	-0.149
SLC5A3	Transport	-0.354	0.210
SLC5A4	Transport	1.235	1.147
SLC5A9	Transport	0.465	0.320

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Sheraj, Ilir
Nationality: Albanian
Date and Place of Birth: 11 January 1989, Sarande
Phone: +90 5398472632
email: sherajilir@gmail.com

EDUCATION

Degree	Institution	Year of Graduation
BSc	METU Biology	2011
MS	METU Biology	2014

FOREIGN LANGUAGES

Fluent English, Intermediate Turkish and German

PUBLICATIONS

1. Tunçer,S., Gurbanov, R., **Sheraj, I**., Solel, E., Esenturk, O., Banerjee, S. (2018). Low dose dimethyl sulfoxide driven gross molecular changes have the potential to interfere with various cellular processes. *Scientific reports* 8 (1), 14828.

2. Torun,A., Enayat,S., **Sheraj,I**., Tunçer,S., Ülgen, DH., Banerjee, S. (2019). Butyrate mediated regulation of RNA binding proteins in the post-transcriptional regulation of inflammatory gene expression. *Cellular Signaling* 64, 109410.

3. Tunçer,S., Sade-Memişoğlu,A., Keşküş, AG., **Sheraj,I**., Güner,G., Akyol,A., Banerjee, S. (2019). Enhanced expression of HNF4α during intestinal epithelial differentiation is involved in the activation of ER stress. *The FEBS Journal.* doi: 10.1111/febs.15152.

4. Demirkol, SC., Seza, EG., **Sheraj, I**., Gömçeli, I., Turhan, N., Carberry, S., Prehn, JHM., Güre, AO., Banerjee, S. (2020). Evaluation of an aldo-keto reductase gene signature with prognostic significance in colon cancer via activation of epithelial to mesenchymal transition and the p70S6K pathway. *Carcinogenesis*, 41, doi.org/10.1093/carci/bgaa072.

5. **Sheraj, I**., Guray, TN., Banerjee, S. (2021). A pan-cancer transcriptomic study showing tumor specific alterations in central metabolism. *Scientific reports* 11, 13637. doi: 10.1038/s41598-021-93003-3.

PROGRAMMING LANGUAGES

R Programming Language (Advanced)

Python (Advanced)

SQL

Julia

HOBBIES AND INTERESTS

Reading all Kind of Books, in particular philosophy, Extreme Sports, Running or Long walks (10-15 km).