MULTI-OMIC DATA INTEGRATION TO STRATIFY POPULATION IN HEPATOCELLULAR CARCINOMA

A THESIS SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAIʻI AT MANŌA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN

MOLECULAR BIOSCIENCE AND BIOENGINEERING

July 2016

By

Liangqun Lu

Thesis Committee:

Dr. Lana Garmire (Chair)

Dr. Maarit Tiirikainen

Dr. Herbert Yu

Acknowledgements

First of all, I would like to express my gratitude to my advisor Dr. Lana Garmire for the support of my study. Her guidance helped me during my research and thesis writing.

Second, I would like to thank my committee members Dr. Maarit Tiirikainen and Dr. Herbert Yu for their advice and comments on my thesis project.

Last but not the least, I thank my fellow labmates in Dr. Lana Garmire's group: Travers Ching, Kumardeep Chaudhary, Sijia Huang, Xun Zhu, Olivier Poirion and Jacob Fennick. I looked for help from them whenever I had problems. And the communication with them always provided me different perspectives. It was a pleasure to work with them. In addition, I would like to thank my friend Runmin Wei for his generous help during my study in Hawaii.

Table of Contents

Index of Figures

Index of Tables

Chapter 1. Introduction

1.1 Hepatocellular carcinoma (HCC) epidemiology

Liver cancer is the malignancy that starts in the liver with late-stage symptoms, including weight loss, loss of appetite and yellowing of the skin and eyes. According to world health organization data, liver cancer is the sixth most common cancer and is the second leading cause of cancer death. The recent statistics on liver cancer showed that 782,500 (554,400 males and 228,100 females) estimated new cases and 745,500 (521,000 males and 224,500 females) deaths happened worldwide (Torre et al. 2015). Liver cancer is more prevalent in developing countries than developed countries. Eastern Asia, South-Eastern Asia, Northern Africa and Western Africa are among those nations with the highest incidence rates. In particular, about 50% of new cases and deaths occurred in China (Torre et al. 2015). Hepatocellular carcinoma (HCC), is the most dominant type of liver cancer, accounting for approximately 80% of the cases, while cancer of the bile duct (cholangiocarcinoma and cholangiocellular cystadenocarcinoma) are only approximately 6% of all liver cancer cases.

In the United States, liver cancer is the fifth most common cause of cancer deaths in males while the ninth most common cause of cancer deaths in females, from National Cancer Institute (NIH) statistics in 2015. An estimated 35,660 adults (25,510 men and 10,150 women) were diagnosed with primary liver cancer, and an estimated 24,550 deaths (17,030 men and 7,520 women) occurred in 2015(Siegel, Miller, and Jemal 2015). Furthermore, 39230 new cases and 27170 deaths are expected in 2016, a slight increase as compared to the previous year.

In clinics, tumor TNM stage (tumor size, lymph nodes and cancer cell metastasis) ranging from I- IV is used to describe primary liver cancer (Figure 1). Stage I indicates a single tumor without spreading to the blood vessels and lymph nodes. Stage II means tumor has grown into blood vessels. Stage III is advanced and has three subgroups, 3A, 3B and 3C. Stage 3A has more than one tumor, as least one of which is larger than 5 cm. Stage 3B has invaded into blood vessels, while stage 3C has grown into nearby organ, such as pancreas. Stage IV is the most advanced

stage and has 4A and 4B subgroups. Stage 4A tumors metastasize into blood vessels and regional lymph nodes, and Stage 4B tumors into other organs. In summary, stage I and II tumors are localized, stage III has spread to nearby organs while stage IV has metastasized. Cancer staging at diagnosis help to determine appropriate treatment strategies and is associated with the length of survival. From the NIH statistics, the five-year survival for localized patients is 30.9%, 10.9% for regional patients and 3.1% for distant patients. Overall, the overall five-year survival rate is about 17%. Another parameter, histologic grade, has been used to describe the degree of cell differentiation: poorly and undifferentiated (grade I), moderate (grade II), or well differentiated (grade III). Similar to stage, grade also influences the length of survival.

Additionally, HCC incidence rate varies among races and age groups. The East Asian population suffering from HCC, both females and males, are twice as likely to develop liver cancer compared to Caucasian or African American populations. This could be attributed to the difference in major risk exposures or genomic loci (El–Serag and Rudolph 2007). Gender disparity is also obvious, in that males have two to four times higher incidence rates than females. This may relate to male-specific behavior (eg. higher chances of alcohol consumption). It may also be associated with gender-specific sex hormone differences. For instance, the high level of prolactin (PRL), an estrogen responsive hormone in females, reduces HCC incidence by restricting innate immune activation (Hartwell et al. 2014; Seton-Rogers 2014). With regard to age, the rate of HCC peak is 65-70 for females and 60-65 for males (El–Serag and Rudolph 2007)(Siegel, Miller, and Jemal 2015).

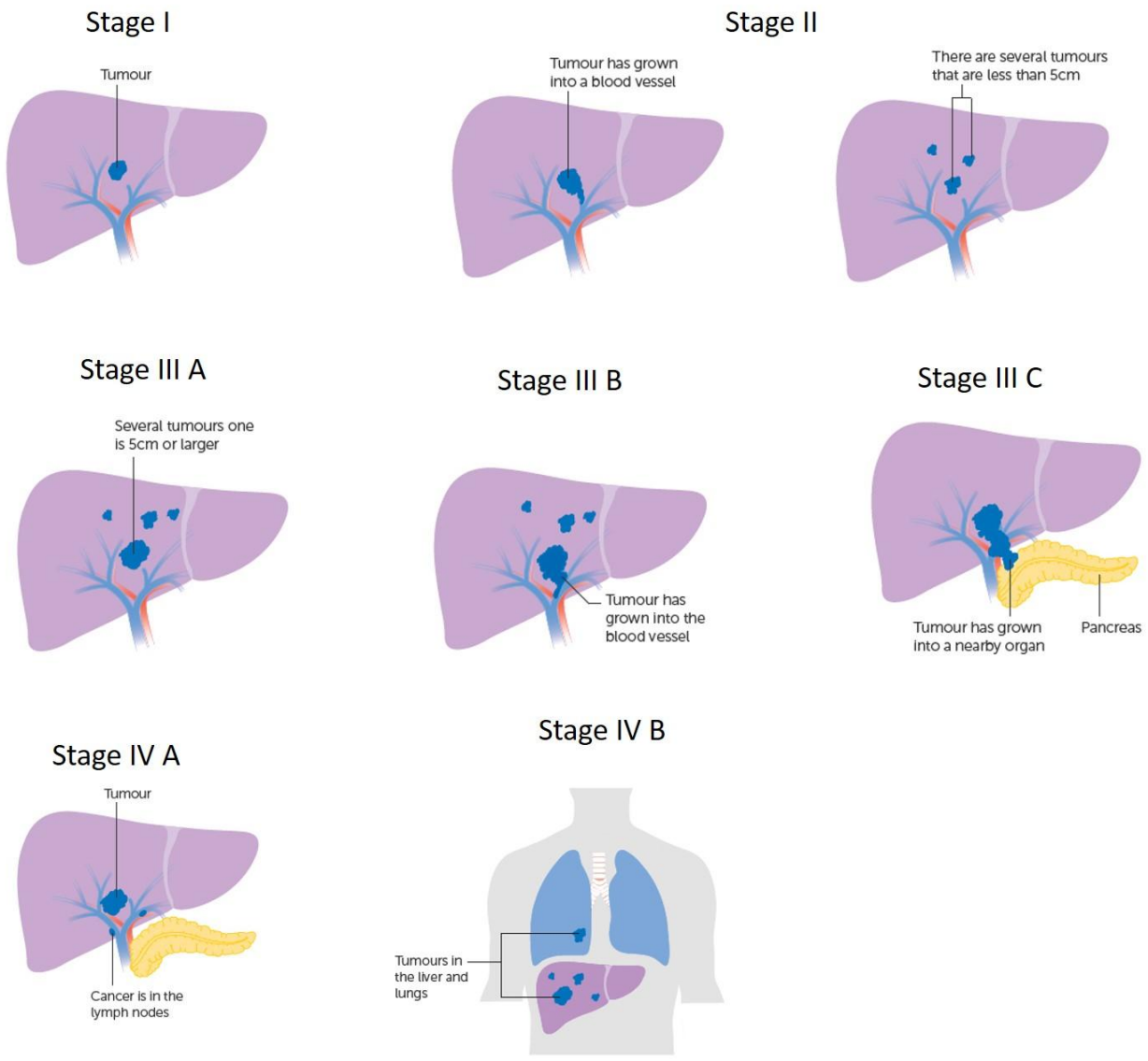

Figure 1: Liver cancer clinical stages, from the Cancer Research UK

Various risk factors for the HCC development have been well studied, including cirrhosis, hepatitis B infection, hepatitis C infection, alcohol abuse, metabolic disorder, obesity and environmental toxic intake. HCC often occurs in patients with liver fibrosis or cirrhosis and rarely develops in the healthy liver. About 70-80% of HCC patients have underlying cirrhosis, a chronic liver damage caused by the inflammation and fibrosis. (European Association for The Study Of The Liver & European Organisation For Research And Treatment Of Cancer, 2012)

(Figure 2) The major risks vary among geographical locations. HBV infection is the major risk for HCC cases in East Asian countries (Laursen 2014). Hepatitis B, a DNA virus, is able to integrate DNA into the host genome and express the transactivator proteins, such as HBxAg, which activate signaling proteins (such as NF-kB, PI3K), upregulate the associated signaling pathways and proliferative processes, and increase the development of HCC (Ayub, Ashfaq, and Haque 2013). HCV is the leading cause of HCC in North America, Europe and Japan. Hepatitis C, an RNA virus, triggers the inflammatory response resulting in increased proliferation and cirrhosis through fibrogenesis (Bühler and Bartenschlager 2012). Aflatoxin B1 is a mutagenic toxin, produced as a secondary metabolite by the fungus *Aspergillus flavus*. Aflatoxin B1 works as a cofactor with HBV to result in HCC in Africa (Llovet et al., 2016). Alcohol-related HCC mostly occurs in the population with low virus infection rates, such as United States and Europe. The long-term abuse of alcohol causes the induction of the CYP2E1 enzyme, which results in the acceleration of hepatic acetaldehyde production, reactive oxygen species (ROS), and hepatic oxidative stress in HCC progression (Testino, Leone, and Borro 2014).

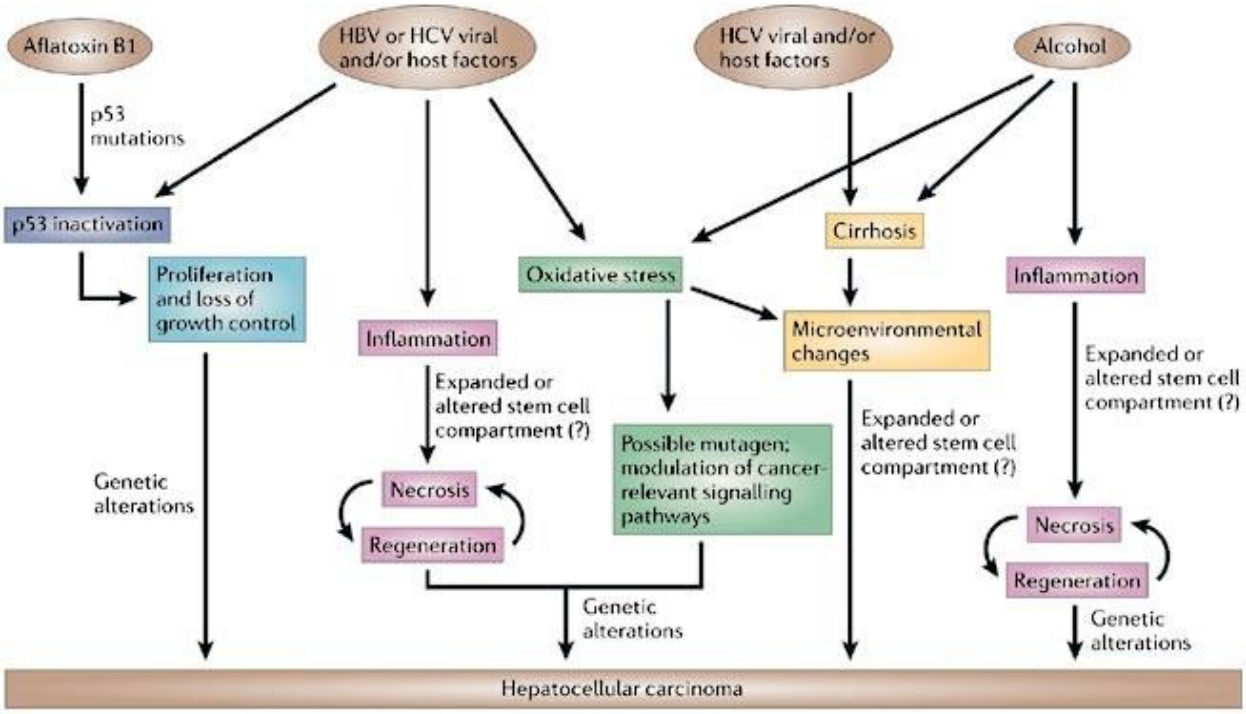

Figure 2: Risk factors and mechanisms of hepatocarcinogenesis, from the review (Farazi & DePinho, 2006)

1.2 HCC etiology

The observation studies from HCC epidemiology increases the urgency of elucidating the mechanism of HCC initiation and progression. Generally, when various risk factors (HBV, HCV, alcohol, aflatoxin B1) injure hepatocyte, there is necrosis followed by the proliferation in hepatocyte. Continuous injuries between risk factors and hepatocyte response foster chronic liver disease, and furthermore cultivate liver cirrhosis, where abnormal nodules are observed. Subsequently, with the increasing genetic and genomic alterations, HCC is eventually progressed from several steps including hyperplastic nodule and dysplastic nodule. (Figure 3)

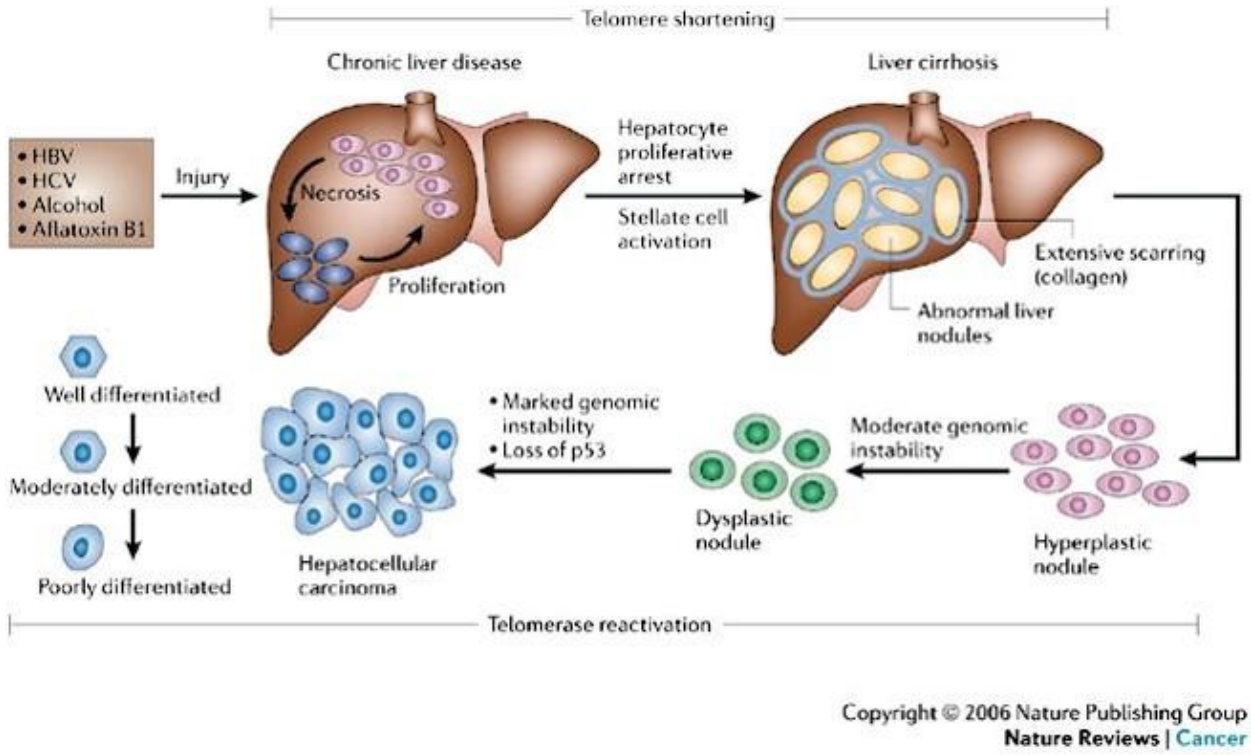

Figure 3: HCC etiology, from (Farazi & DePinho, 2006)

Processes including telomere shortening, loss and/or mutation of p53 and genomic instability all play roles in hepatocarcinogenesis. Telomere shortening characterizes chronic hyperproliferative liver disease, which is assumed to be associated with hepatocyte turnover and contribute to genomic instability in the initiation and progression of HCC. Telomerase Reverse Transcriptase (TERT) activation and short telomere are necessary for malignant progression. It is widely documented that p53 deficiency is associated with the development of various cancers, including HCC. The HBV- and HCV-related HCCs have been associated with a high frequency of p53 mutation. Additionally, loss of p53 is common in HCC initiation and development, which is assumed to facilitate continued proliferative potential, activate DNA-damage signaling and increase genome instability. Genomic instability is a common feature in HCC that may be attributed to telomere shortening, chromosome segregation defects and alterations in the DNA-damage-response pathways.

1.3 Genomic landscape research on HCC

The development of high-throughput technologies has sped up our understanding of genetics, epigenetics, mRNA and non-coding RNA transcriptomics. In recent years, a large number of omic research have been done to detail the mechanism of molecular pathogenesis of HCC. This has significantly contributed to our understanding of the cancer genomics, diagnostics, prognostics, and therapy in an unprecedented way. In particular, DNA-level and RNA-level omic research, including exome sequencing, copy number somatic mutation, DNA methylation, mRNA and miRNA sequencing, accelerate the investigation of HCC initiation and development.

1.3.1 Somatic alterations

It is well accepted that virtually all cancers are the results of accumulated somatic alterations in genome, which lead to tumor proliferation and fitness of adaptation. With the advances in high-throughput technologies, it is common to infer driver alterations from passenger alterations. Driver mutations are defined as mutations that confer a selective growth advantage to the cell while passenger mutations refer to those which do not alter the fitness. It is believed that only a small fractions of all mutations contribute to the initiation and development in cancer (Vogelstein et al. 2013). Whole-genome and whole-exome sequencing provide a comprehensive and high-resolution landscape of somatic genomic alterations in HCC.

Research groups from Japan, China, United States, France and Korea have investigated putative driver genes in HCC, although they have variable sample sizes (M. Li et al. 2011; Schulze et al. 2015; Fujimoto et al. 2012; Ahn et al. 2014; Totoki et al. 2011; Kan et al. 2013) (Table 1). TP53 and CTNNB1 are two most frequently mutated genes in HCC cases. Candidate driver genes related to genome stability have also been discovered, such as ARID1A, ARID2, MLL1-4 (M. Li et al. 2011; Guichard et al. 2012; J. Huang et al. 2012; Fujimoto et al. 2012; Cleary et al. 2013). Other important candidate driver genes include RB1 involved in cell cycle control, AXIN1 in Wnt signaling pathway, NFE2L2 in oxidative stress, TSC1 in MAPK signaling pathways. In addition, the public database COSMIC (Catalogue of somatic mutations in cancers) collects

HCC mutated genes, and frequently mutated genes are TP53, CTNNB1 inferred from 664 HCC samples. The frequently mutated genes from HCC cell line studies are TP53, AKAP9 (Forbes et al. 2015).

Table 1: Candidate driver gene identification in HCC

Study	Group	Technique	Sample size	Candidate onco-event
Li et al.(M. Li et al. 2011)	China	WES	10	ARID2 (inactivation)
Guichard (Guichard et al. 2012)	France	WES	24	ARID1A, RPS6KA3, NFE2L2, IRF2 (new)
J. Huang et al.(J. Huang et al. 2012)	China	WES	110	ARID1A
Cleary et al.(Cleary et al. 2013)	US	WES	87	MLL
Totoki et al.(Totoki et al. 2011)	Japan	WGS	10	TSC1 (nonsense substitution)
Fujimoto et al.(Fujimoto et al. 2012)	Japan	WGS	27	ARID1A, ARID1B, ARID2, KMT2A, MLL3
Kan et al.(Kan et al. 2013)	US	WGS	88	Wnt pathway, JAK/STAT pathway
Ahn et al.(Ahn et al. 2014)	Korea	WES	231	TP53, CTNNB1, AXIN1, RPS6KA3, RB1
Schulze et al.(Schulze et al. 2015)	France	WES	243	CTNNB1; TP53; AXIN1

* WGS: whole genome sequencing, WES: whole exome sequencing

As another type of of DNA-level alterations, copy number variation (CNV) plays important roles in carcinogenesis and progression, including HCC. The analysis on copy number alteration has identified various loci in cancer development (Shibata and Aburatani 2014). Shibata and Aburatani summarized recurrent copy number variations in HCC (Table 2). The amplification of 8q24 (MYC), 1q32.1 (MDM4), 20q13.33 (EEF1A2) showed increased HCC progression

(Schlaeger et al. 2008). In another cohort of 286 HCC samples, BCL9 (1q21.1) and MTDH (8q22.1) were identified as novel amplified oncogenes (K. Wang et al. 2013).

Table 2: Copy number variance events in HCC

Locus	CAN	Gene name	Function
1p36.11	deletion	CDKN2C	Cell cycle
1p36.11	deletion	ARID1A	Chromatin remodelling
1p36.33	deletion	TNFRSF14	Immune response
1q21.1	amplification	BCL9	WNT pathway
1q21.2	amplification	ARNT	Xenobiotics metabolism
1q25.2	amplification	ABL2	Proliferation
1q32.1	amplification	MDM4	p53 pathway
6q26	deletion	TNFAIP3	NF-κB pathway
7q31.2	amplification	MET	Proliferation
8p11.2	deletion	PROSC	Unknown
8p21.2	deletion	SH2D4A	Proliferation
8p21.3	deletion	SORBS3	Migration
8p21.3	deletion	WRN	DNA repair
8p22	deletion	DLC1	Small GTPase
8p23.2	deletion	CSMD1	Immune response
8q13.1	amplification	COPS5	Proteolysis
8q22.1	amplification	MTDH	Metastasis
8q22.2	amplification	COX6C	Mitochondria
8q24.21	amplification	MYC	Proliferation

9p21.3	deletion	CDKN2A	Cell cycle
9p21.3	deletion	CDKN2B	Cell cycle
10q23.31	deletion	PTEN	Proliferation
11q13.2	amplification	CCND1	Proliferation
11q13.2	amplification	FGF19	WNT pathway
13q11	deletion	XPO4	Nuclear export
13q13.1	deletion	BRCA2	DNA repair
13q14.3	deletion	RB1	Cell cycle
13q31.1	deletion	SPRY2	Proliferation
17q23.1	amplification	RPS6KB1	Proliferation
18q21.31	deletion	SMAD4	TGF-β signalling
20q13.33	amplification	EEF1A2	Translation

Presumably, driver genes contribute to tumor initiation, development, metastasis and drug resistance. At the pathway level, these candidate driver alterations arise from five pathways (Zucman-Rossi et al. 2015)(Shibata and Aburatani 2014). Wnt signaling pathway (CTNNB1 and AXIN1) is frequently activated as mutant CTNNB1 activates the beta-catenin pathway. AXIN1 is usually inactivated after mutation (Totoki et al. 2014). Genes in cell cycle pathway (TP53, RB1 and CDKN2A) are frequently mutated in HCC. The inactivation of TP53 and RB1 and the deletion of CDKN2A are common, which are associated with a poor prognosis and could contribute to a more aggressive phenotype. Chromatin remodeling complexes and epigenetic regulators (ARID1A and ARID2) are also frequently altered. AKT–mTOR– MAPK signaling (TSC1 and TSC2) are frequently activated in HCC. Oxidative stress pathway is constitutively activated in HCC due to mutations that activate nuclear factor elytroid 2-related factors 2 (NFE2L2) or inactivate KEAP1 in HCC cases. In addition, telomere maintenance contributes to the evasion of cellular senescence. In previous studies, TERT is overexpressed in 90% of HCC

due to promoter mutation (60% cases) and gene amplification (5% cases) (Figure 4). Table 3 below shows the major driver genes and their pathways in HCC.

Table 3: HCC progression and somatic alterations, from the Hepatocellular carcinoma Nature Reviews Disease Primers (Llovet et al., 2016)

Pathway(s)	Gene(s)	Alteration	Frequency in HCC	Experimental evidence of "driver" properties
Telomere maintenance	TERT	Promoter mutation	54–60%	Yes
		Amplification	5–6%	
Cell cycle control	TP53	Mutation or deletion	12–48%	Yes
	RB1	Mutation or deletion	3–8%	
	CCND1	Amplification	7%	
	CDKN2A	Mutation or deletion	2–12%	
WNT–β- catenin signalling	CTNNB1	Mutation	11–37%	Yes
	AXIN1	Mutation or deletion	5–15%	NA
Oxidative stress	NFE2L2	Mutation	3–6%	Yes
	KEAP1	Mutation	2–8%	Yes
Epigenetic and chromatin remodelling	ARID1A	Mutation or deletion	4–7%	NA
	ARID2	Mutation	3–18%	NA
	KMT2A (MLL1), KMT2B (MLL4), KMT2C (MLL3) and KMT2D (MLL2)	Mutation	2–6%	NA
AKT–mTOR–	RPS6KA3	Mutation	2–9%	NA

MAPK signalling	TSC1 and TSC2	Mutation or deletion	3–8%	Yes
	PTEN	Mutation or deletion	1–3%	
	FGF3, FGF4 and FGF19	Amplification	4–6%	
	PI3KCA	Mutation	0–2%	

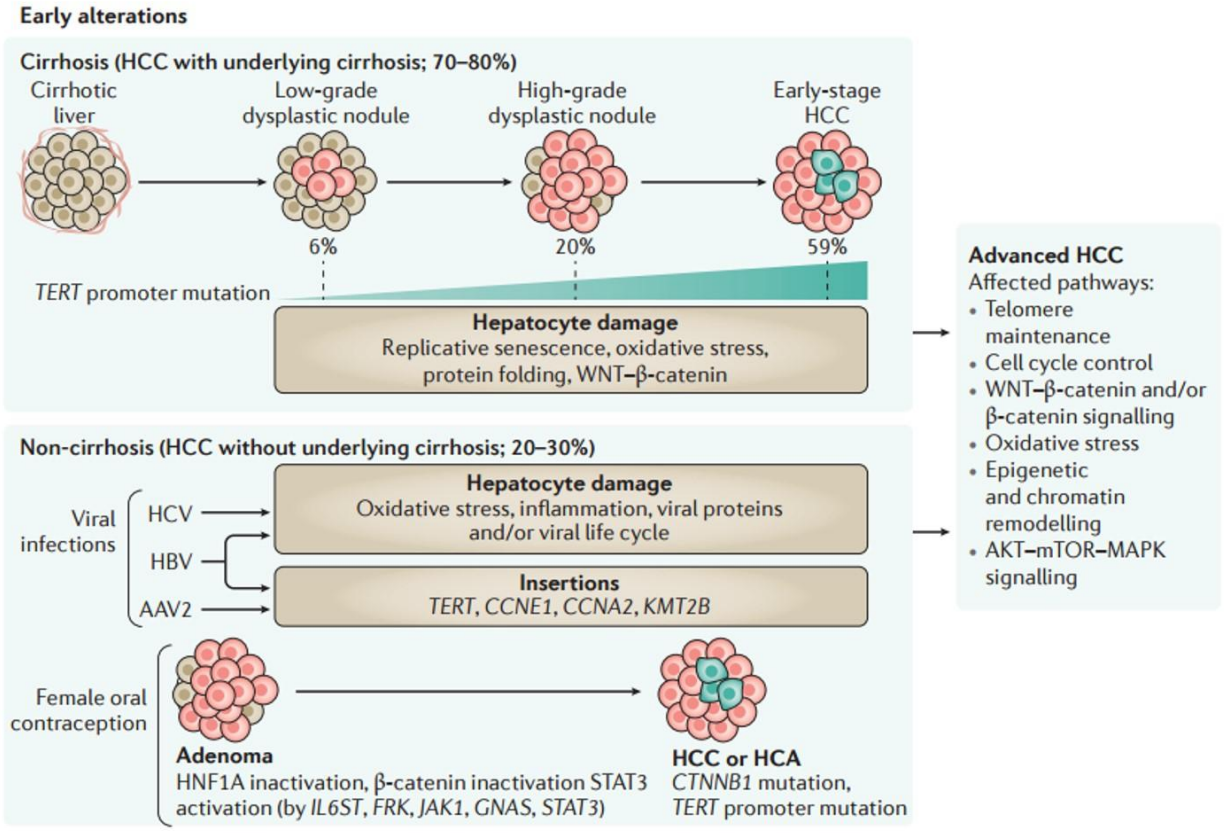

Figure 4: HCC progression and somatic alterations, from the Hepatocellular carcinoma Nature Reviews Disease Primers (Llovet et al., 2016)

1.3.2 DNA methylation

Epigenetic abnormalities play roles in tumorigenesis and tumor development, since they cause aberrant gene expression (Dumitrescu 2009). DNA methylation is a most common study to delineate epigenetic mechanism in human carcinogenesis. DNA methylation is the biological process that methyl groups are added to the nucleotides during or after DNA replication. When located in CpG islands, DNA methylation regulates gene expression and is necessary for cell differentiation and participate in tumorigenesis. Global loss of DNA methylation has been discovered in cancers, suggesting epigenetic reprogramming. Global DNA hypomethylation and promoter gene CpG hypermethylation are dominant for HCC development (Calvisi et al. 2007).

The hypermethylation of multiple genes, Ras-association domain family 1, isoform A (RASSF1A), BLU and fragile histidine triad (FHIT) commonly occurs in early stage of HCC cases. On the other hand, Retinol Binding Protein 1 (CRBP1) methylation is involved in late stage HCC (Zhang et al. 2013). DNA methyltransferases (DNMTs) are important players in hepatocarcinogenesis. DNMT3a and DNMT3b have increased expression from chronic liver diseases to HCC. DNMT3a was assumed to improve cell proliferation by modulating phosphatase and tensin homolog (PTEN) expression (Zhao et al. 2010). Similarly, DNMT3b targeted metastasis suppressor 1 (MTSS1), a tumor suppressor in HCC (Fan et al. 2012). DNA methylation changes are also detected in genes involved in cell cycle gene. Ubiquitin carboxyl-terminal hydrolase L1 (UCHL1) and fructose-1,6-bisphosphatase-1 (FBP1) were identified hypermethylated in HCC cell lines, which led to low expression and inhibited cell proliferation through G2/M phase cell cycle arrest (Chen et al. 2011).

1.3.3 Transcriptomics

Microarray-based and RNA-sequencing technologies have been applied to identify molecular and genomic mechanisms in HCC development. Several studies have been done on gene expression in an attempt to extract featuring expression patterns and distinguish HCC from normal tissues, correlate with risk factors and clinical phenotypes, and determinate novel subtypes and regulatory network in treatment (Maass et al. 2010; Wurmbach et al. 2007; Tsai et

al. 2006). Huang et al (Q. Huang et al. 2011) shows that 1378 genes are differentially expressed in HCC (HBV+). Downstream enrichment analysis of these genes showed a significant correlation with chromosome location on 8q21.3–24.3. Another group (RNA-seq and CNV) indicated that chromosome 8q was the most predictive of overall survival and that 22/50 potential driver genes were located in this region (Woo et al. 2009).

1.3.4 MiRNA expression

MiRNA is a class of 19-23 nt non-coding RNAs that regulate genome-wide gene expression by either destabilizing targeted mRNAs and/or inhibiting their protein translation. With more and more miRNAs identified and functionally annotated, it has been increasingly recognized that miRNAs play an important role in human carcinogenesis (Bartel 2004; Meng et al. 2007). Morishita and Masaki summarized the up-regulated and down-regulated miRNAs in HCC. MiR-21, miR-221 and miR-222 were up-regulated in HCC. And miR-122a, miR-145, miR-199a and miR-223 were down-regulated (Morishita and Masaki 2015). Other public databases provide information about miRNA expression change in HCC. For example, miR-1301, hsa-miR-155, hsa-miR-21, hsa-miR-221, hsa-miR-27a and hsa-miR-525-3p are up-regulated in liver cancer, from miRCancer database (Xie et al. 2013). In the diagnosis, plasma miR-21 level in HCC patients was significantly higher than patients with chronic hepatitis and healthy volunteers (Tomimaru et al. 2012).

1.4 Molecular classification in HCC

Through NGS applications in HCC study, it is possible to subgroup samples that share similar genetic features or clinical characteristics. These classifications are potentially helpful to classify the key responders in clinical trials and guide therapeutics, however the task to build such classes is demanding (Goossens, Sun, and Hoshida 2015). In an investigation of genotype-phenotype correlations, Boyault et al. clustered 57 tumor samples into 6 groups (G1-G6) based on gene expression profiling from HG-U133A array with 6712 probe sets(Boyault et al. 2007). Another meta-analysis study analyzed 9 independent datasets on a total of 603 HCC samples and revealed

three subclasses (S1, S2 and S3) from the transcriptional level (Hoshida et al. 2009). The analysis of subgroup characteristics shows that S3 is less aggressive with better survival and enriched with CTNNB1 mutation, whereas S1 and S2 have more common TP53 mutations (Figure 6).

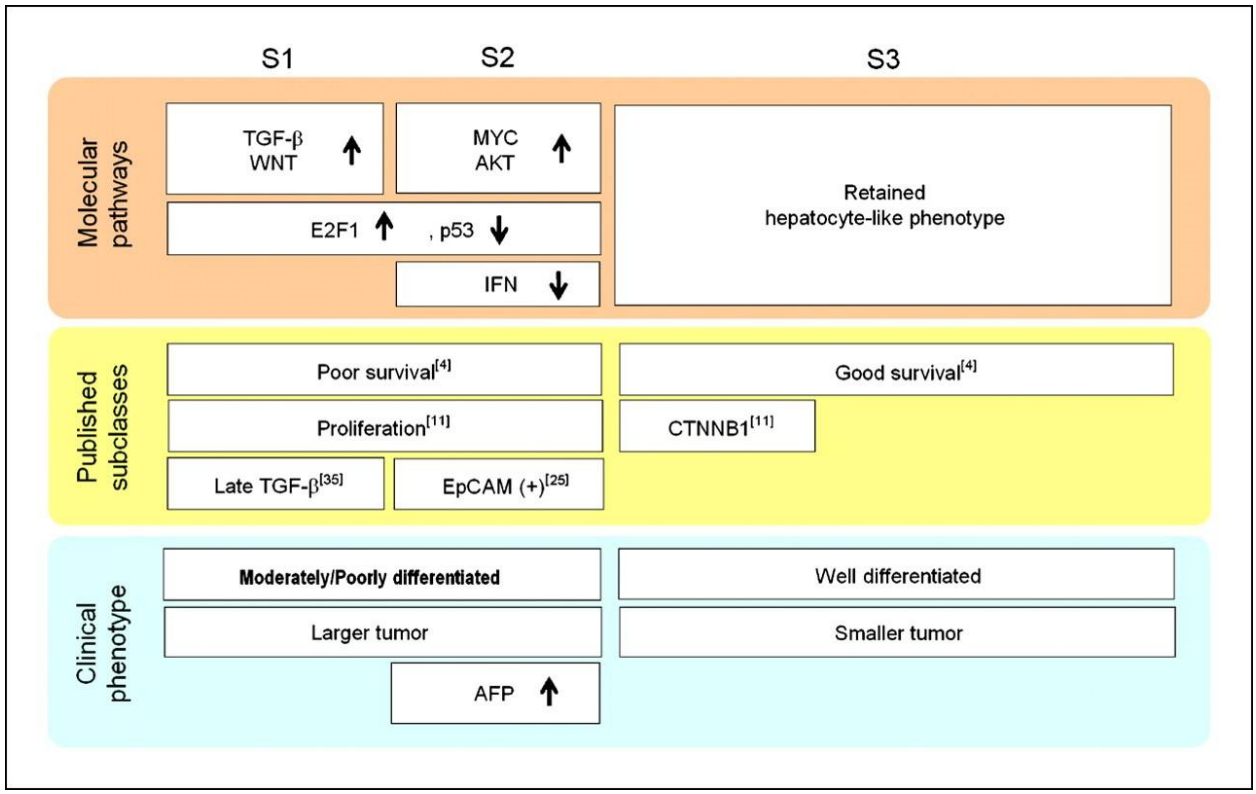

Figure 5: Schematic summary of the characteristics of HCC subclasses, from the article(Hoshida et al., 2009).

1.5 Diagnostic and prognostic molecular markers

The patterns and expression signatures are promising to function as biomarkers for HCC diagnostics and prognostics. Early diagnosis is key for the survival of HCC patients. A series of biomarker research on HCC explored the potentials across biological molecules, including protein antigens, enzymes, cytokines and non-coding RNAs (Zucman-Rossi et al. 2015). Moreover, through the integration of molecular markers in the existing clinical staging system

(tumor size, number of nodules, tumor stage and grade), it is believed to improve accuracy of diagnostic and prognostic processes.

As one of the earliest and most popular serum markers, AFP (alpha-fetoprotein) and AFP glycoforms (AFP-L3) have an elevated expression in HCC and cirrhotic patients. They are used as a diagnostic marker and a prognostic marker for tumor development after the patient treatment. In addition, the overexpression of proteins in HCC tissues, HSP (heat shock proteins) family members HSP70 and HSP27, glypican-3 (GPC3), GP73 (Golgi protein 73) and TAG-72 (Tumor-associated glycoprotein 72) are assumed to promote HCC tumor growth and progression. Enzymes DCP (Des-γ-carboxyprothrombin), GGT (γ-glutamyl transferase) and AFU (α-l-fucosidase) were detected with high expression in the serum of HCC patients and mark the disease stages. The cytokines TGF-β1 (Transforming growth factor-β1) and VEGF (endothelial growth factor) have high expression in the serum and tissue of HCC patients individually and work as indicators for HCC patients. Although all these molecules were concluded from large cohorts of studies, the clinical validation proves that single molecule cannot be an indicator the status of patients (Y.-J. Zhao, Ju, & Li, 2013). New classes of non-coding RNAs, such as miRNAs are also proposed as candidate biomarkers in in HCC. miR-500, miR-29 and miR-122 were identified down-regulated in HCC patients while miR-21 was up-regulated in the serum. miR-21 was assumed to repress the expression of PENT and PENT-related pathways and promote the tumor growth.

1.6 Objectives in this study

Although significantly mutated genes have been identified in a series of cohort studies in HCC, the association with gene expression and clinical characteristics has not been investigated. What is more, building molecular subgroups will help to guide precise targeted therapeutics. Such classification research considering the comprehensive effect from different level information is lacking investigation on HCC. Our goal in this study is to associate putative driver genes with clinical characteristics and gene expression, and build integrative clusters on HCC samples. We

have utilized TCGA datasets in this study, with additional validation cohorts. Specifically, we have two aims:

Specific aim 1: Clinical and transcriptomics associations of putative driver mutations in hepatocellular carcinoma

Specific aim 2: Multi-omic data integration to determine subtypes and build predictive models on new patients.

Chapter 2. Materials and Methods

2.1 HCC samples and processing

We explored the liver hepatocellular carcinoma data from the TCGA portal (https://tcga-data.nci.nih.gov/tcga/) on Sep. 25, 2015. We used R package TCGA-assembler (v1.0.3) (Zhu, Qiu, and Ji 2014) and downloaded the omic data, somatic mutations (BCM__Mixed_DNASeq_curated; Level 2) maf file, RNA-seq data (UNC IlluminaHiSeq_RNASeqV2; Level 3), miRNA data (BCGSC IlluminaHiSeq_miRNASeq; Level 3), DNA methylation data (JHU-USC HumanMethylation450; Level 3), Copy number variation data (BI Genome_Wide_SNP_6; Level 3) and the clinical information (Table 5 and Table 6). Specifically, a total of 198 samples have paired tumor and normal adjacent tissue whole exome-seq data. 360 tumor tissues and 39 tumor-adjacent normal tissues are available as all four datasets, RNA-seq, miRNA-seq, DNA methylation and Copy number variation. For the DNA methylation, we mapped CpG islands within 1500 bp transcription start site (TSS), both hypermethylation and hypomethylation, to nearby genes and obtained the combined effect. Also for copy number variation, we used segment values from SNP-calling file on hg19, which includes germline mutation and somatic mutation. In dealing with the missing values, three steps were processed. First, the biological features (for example, genes) were removed if more than 20% were missing across all patients. Similarly, the samples were removed if more than 20% features missing. Second, we used *impute* function from R impute package (Xiang et al. 2008), which is based on K nearest neighbor, to fill out the missing values. Last, we removed genes with 0 across all samples.

Table 4: TCGA Liver hepatocellular carcinoma (LIHC) data

Data types	Platforms	Samples
Exome-sequencing	Illumina Genome Analyzer DNA Sequencing	198 (T) 198 (N)
Copy-number Variation	Affymetrix Genome-Wide Human SNP Array 6.0	371 (T) 83 (N)
RNA-seq	Illumina HiSeq 2000 RNA Sequencing Version 2 analysis	371 (T) 50 (N)
DNA Methylation	Illumina Infinium Human DNA Methylation 450	377 (T) 50 (N)
miRNA-seq	Illumina HiSeq 2000 miRNA Sequencing	372 (T) 51 (N)
Clinical		377 cases

* T: tumor samples; N: normal samples (solid tissue)

TCGA provides diverse clinical record information for patients. Therefore, we list the clinical characteristics of HCC samples below. Of 377 cases, 325 patients have associated available survival status, including overall survival and disease-free survival. The gender, stage, grade and risk factors are revealed in Table.

Table 5: Clinical Characteristics of HCC Samples

Characteristics	Patients 377 (100%)
Age	59.5 +- 13.5 (99.7%)
Gender	
Female	122 (32.4%)
Male	255 (67.6%)
Stage	
Stage I	175 (46.42%)
Stage II	88 (23.34%)
Stage III	86 (22.81%)
Stage IV	6 (1.59%)
Grade	
Grade 1	55 (14.59%)

Grade 2	180 (47.75%)
Grade 3	124 (32.89%)
Grade 4	13 (3.45%)
Race	
White	187 (49.60%)
Asian	161 (42.71%)
African American	17 (4.51%)
American Indian or Alaska Native	2 (0.53%)
Risk factors	
HBV	78 (20.69%)
HCV	32 (8.49%)
Alcohol	69 (18.30%)
Alcohol & HBV	20 (5.31%)
Alcohol & HCV	14 (3.71%)
Fatty Liver Disease	11 (2.92%)
No Primary risk	93 (24.67%)
Others	60 (15.93%)
Survival status	
Overall Survival	325 (98 censored) (86.2%)
Disease free survival	298 (142 censored) (79.0%)

2.2 Putative driver gene calculation

MutSig is a computational tool developed to process mutations detected in DNA-sequencing and identify significantly mutated genes, also meaning putative driver genes, given the background mutation rate (BMR). In DNA replication, base mutation rate is 1e-9 each time. This normal error rate from replication holds diversity and adaptability. In cancer evolutionary biology, these selected genes are hypothesized to be positive selective to tumorigenesis (the International Cancer Genome Consortium Mutation Pathways and Consequences Subgroup of the Bioinformatics Analyses Working Group 2013), compared to the neural selection of the non-significant mutated genes, as all genes have the same environmental pressures. In the estimation of gene background mutation, MutSig considers the gene-specific mutation, gene-context and patient-context mutation, gene covariance with neighboring genes, as well as mutation effect from DNA replication and chromatin status. In the estimation of mutation significance, a statistical model explores the background mutation rate, mutation clustering in gene hotspots and

gene conservation and detect genes mutated often than expected by chance. Although many computational tools are developed to identify genetic variants in cancer genomes, MutSig is used widely due to the precise model in the sophisticated processes of gene- and patient- specific background mutation rate and mutational significance. MutSigCV is a new version of a series of MutSig versions, which was used to infer the putative driver genes in this study.

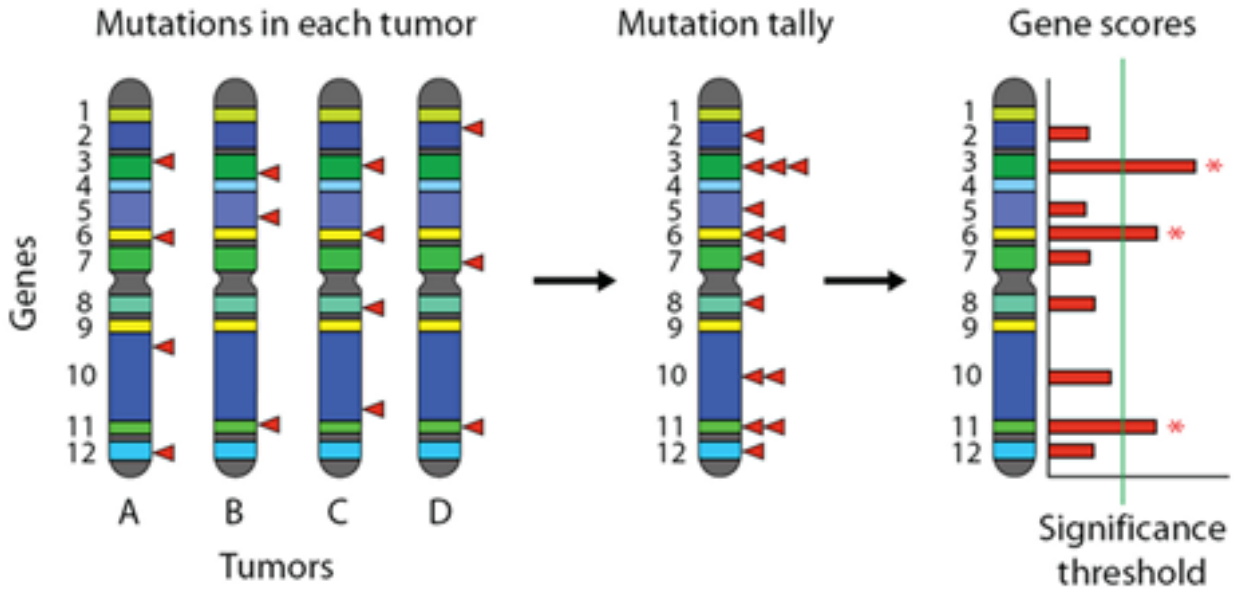

Figure 6: the illustration of the method MutSig, from (Lawrence et al., 2013) (Lawrence et al., 2013)

For the whole exome sequencing data, we calculated the gene mutation statistical significance using MutSigCV tool through the Firehose project (http://firebrowse.org/). MutSigCV is an advanced version of MutSig tool, which estimates genes with positive selection during tumorigenesis through calculating background random mutation rates, clustering within-gene mutations and mutated position conservation (Lawrence et al. 2013). We used the curated somatic mutation file (TCGA BCM__Mixed_DNASeq_curated.maf) as the input exome-seq data for MutSigCV, and obtained the gene mutation computation results on Oct. 06, 2015. We selected the putative driver genes by the recommended threshold q = 0.1, as was done by others (Y. Y. Li et al. 2015; Lawrence et al. 2013; Cancer Genome Atlas Network 2015).

2.3 Modeling between somatic mutation and expression

Specifically, we made a binary (1, 0) table to indicate the mutation status of putative driver genes in all samples. A value of 1 means the existence of at least one variant with biological impacts within the gene body, in the categories of nonsense, missense, frameshift mutation, splice site, transcription starting site and nonstop mutation. Otherwise, 0 was assigned to the gene. We used the function *voom* (*limma* package in R) to pre-process RNA-seq and miRNA-seq prior to linear modeling (Law, Chen, Shi, & Smyth, 2014), then fit the linear models by generalized least squares. These linear models consider the effects of multiple putative driver genes (predictors) on expression values of individual genes (responses) in the following way:

$$y_{sg} = \sum_{i=1} \beta_i X_{is} + \beta_{0g} + \epsilon$$

Where y_{sg} is the expression value of gene/ miRNA, g of sample s. β_{0g} is that baseline expression of g. X_{is} indicates the mutation status of the putative driver gene in sample s. β_i indicates coefficients of putative driver genes. We performed multiple hypothesis tests (MHT) on the significance values of the coefficients β_i using Benjamini–Hochberg adjustment. We also permutated the mutation status of each driver gene in random across samples and built a random linear model, which was used to verify the accuracy. Then we conducted the "hierarchical" multiple testing procedure and t-statistics, and detected the individual coefficients significantly different from zero as gene sets of up-regulated and down-regulated genes.

2.4 Determination of mutual exclusivity and co-occurrence

For each pair of putative driver genes, we made a 2 by 2 contingency table on their mutation occurrences and determined their association based on Fisher's exact test p-value <0.05. Upon significant p-value, a pair is called "co-occurrence" if the log odds ratio was more than 0, or "exclusive" otherwise. For multiple genes k (or a gene set), we also used the Dendrix algorithm (Vandin, Upfal, and Raphael 2012) to identify exclusive mutations across all samples. We used gene sets k= 3,4,5 and calculated their maximum weight with consideration of mutated genes and

samples. We run 100,000 iterations using Markov chain Monte Carlo approach (MCMC) to calculate empirical p-values for the top gene sets with maximum weight.

2.5 Survival analysis on putative driver mutations

We used a Cox proportional hazards (Cox-PH) model (Cox 1972) implemented in R *survival* package for the relapse free survival (RFS) analysis of putative driver genes. For each putative driver gene, we built a univariate CoxPH model and calculated the hazard ratio (HR) of mutation relative to the wild type. We also conducted multivariable Cox-PH model to fit the overall effect of all 13 driver genes on RFS. For this we used R *glmnet* package, since it enables penalization through ridge regression. We did 10 fold cross-validation to optimize the coefficients for each features. In order to evaluate the performance of the survival models, we calculated the concordance index using function *concordance.index* in R *survcomp* package (Schröder et al. 2011), which is based on Harrel's C statistics (Harrell, Lee, and Mark 1996). We divided the samples into high and low risk groups based on median prognosis index (PI) score, the fitted survival values of the Cox-PH model (S. Huang et al. 2014). We plotted the Kaplan-Meier survival curves on the two risk groups and calculated the log-rank p-value of the survival difference between them.

2.6 Data integration using Similarity Network Fusion (SNF)

Similarity Network Fusion (SNF) functions as a computational approach for data integration from different omic data sets (B. Wang et al. 2014). Generally, SNF calculates a sample similarity matrix for omic data sets, such as mRNA expression, DNA methylation and miRNA expression individually. Then SNF integrates these sample similarity matrices iteratively into a comprehensive sample similarity matrix using graph fusions. SNF is assumed to take advantage of common as well as complementary information from different levels of omic datasets. SNF calculates the sample similarity of each data set in the consideration of all genes or miRNAs, which is expected to minimize the feature selection bias, noise and barriers from different data

types. In our molecular classification, we used SNF to integrate four datasets, RNA-seq, miRNA-seq, DNA methylation and copy-number alteration and build robust subgroups.

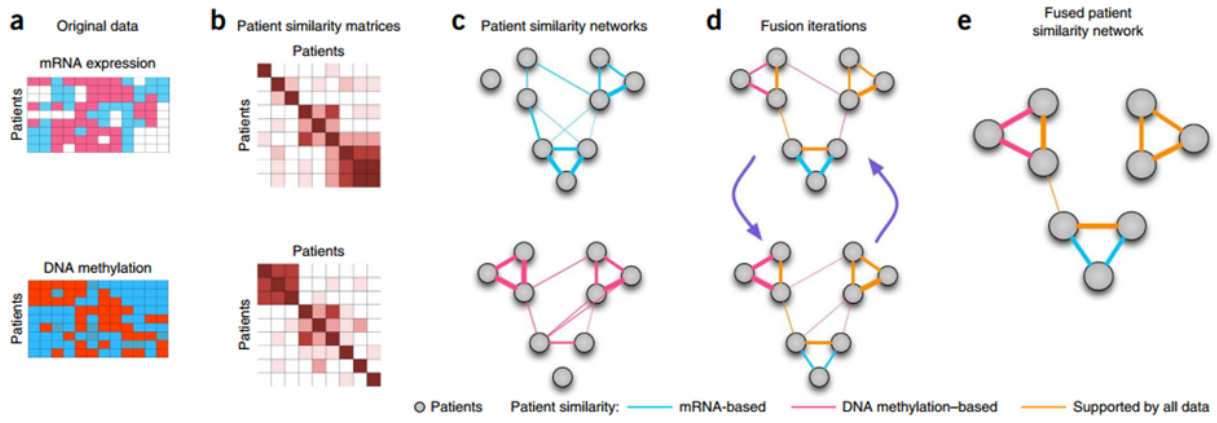

Figure 7: the illustration of the method SNF, from (B. Wang et al., 2014)

For the sample x feature matrix from each omic data, we normalized the matrix by each column and calculated Euclidean distances between all pairs. Then we computed affinity matrices from the genetic distance matrices. For each affinity matrix, we integrated multiple omic information using Similarity Network Fusion (SNF), a computational approach for data integration from different omic data sets (B. Wang et al. 2014), which integrates these sample similarity matrices iteratively into a comprehensive sample similarity matrix using graph fusions.

2.7 Unsupervised clustering on the integrative sample matrix

We used Spectral clustering, one of the unsupervised clustering methods, to build subtypes in the integrative sample similarity matrix. Spectral clustering makes use of the spectrum of the similarity matrix to reduce dimensions and clusters in fewer dimensions. In order to determine the optimal numbers on the HCC samples, we tried different number of clusters from 2 to 8. We then calculated Dunn index (DI), as an internal evaluation to evaluate the separation between clusters by measuring the ratio between inter-cluster and intra-cluster distance. We also calculated the adjusted rand index (ARI) to measure the similarity between clusters. In the

visualization of clustering, we presented principal component analysis plots with group color marked.

2.8 Survival analysis on HCC subtypes

We performed survival analysis using R survival package (Therneau and Grambsch 2000). We fit a Cox-proportional hazards (CoxPH) model between subtypes and clinical outcome, and used likelihood-ratio test to determine the association. We also fit a full CoxPH model of all possible survival related factors, including gender, race, stage, grade, risks and our identified molecular subtypes. We used ANOVA test to measure the significance between models. Furthermore, we plot Kaplan-Meier survival curves for different subtypes.

2.9 Pathway enrichment analysis and network module discovery

We conducted pathway enrichment analysis of the genes impacted by somatic mutations, using using R package clusterProfiler (Yu et al. 2012). We used BH adjusted p value =0.05 threshold to select the over-represented KEGG pathways and presented the network of putative driver genes and pathways in Cytoscape.

2.10 Predictive model on each omic dataset using nearest shrunken-centroid approach

Nearest shrunken centroid method is an advanced version from the nearest centroid method. In the nearest centroid method, the standard centroid for each class is computed through average gene expression for each gene across each class (Tibshirani et al. 2002). In the distance comparison of gene expression of new samples with the centroid of each class, the closest class is assigned to the new samples. In the modification of nearest shrunken centroid method, a threshold is introduced, as defined by the amount of shrunken of class centroids toward the overall centroid for all classes. For instance, a threshold 2.0 means that a centroid of 3.2 would be shrunk to 1.2. After shrunken by the threshold, the distance for new samples with the classes

is calculated and the closest class will be assigned to the new samples. Through the nearest shrunken centroid, it is more accurate as it reduces the possible noise in gene expression. Also it automatically select subset of genes for each class.

In this study, we used the nearest shrunken centroid method to identify subset of genes that best characterize each class, after we obtained the subtypes from SNF method. Assignment based on the expression classifier was concordant with the combined classification in 80% samples, indicate/ demonstrating the efficacy of the approach.

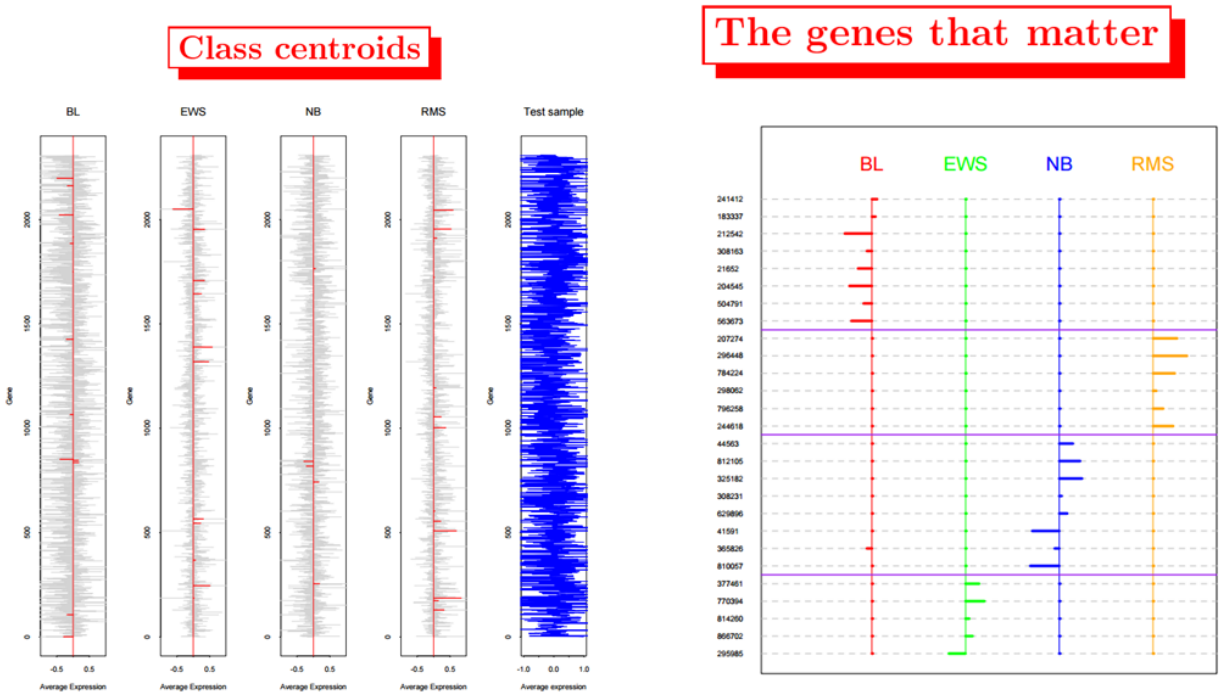

Figure 8: the illustration of the method nearest shrunken centroid, from(Tibshirani et al., 2002)

We took the subtypes from omic data integration as the response classes and divided each omic data into training section (80%) and testing section (20%). We derived the expression signatures of subtypes using the shrunken centroids approach and trained predictive models based on the signatures using the R pamr package (Tibshirani et al. 2002), which summarizes a class centroid for each subtype and assigns new samples to the known subtype with a closest centroid. We

calculated the concordance between one-omic data model and fused omic data on training data, 10 fold cross-validation and testing data.

Chapter 3. Clinical and transcriptomics associations of putative driver mutations in hepatocellular carcinoma

Hepatocellular carcinoma (HCC) is the representative of malignant liver cancer with poor survival. It is widely believed that driver genes contribute to tumor advancement including initiation, progression, metastasis and drug resistance. Although critical driver aberrations have been identified on HCC, the downstream progression of driver genes to tumor development in HCC are not well understood. In this study, we associated somatic mutation status with expression profiles and estimated the expression of genes and miRNAs determined by putative driver genes. Our findings showed that approximately 45% genes and 18 miRNAs have significant relation with putative driver genes. The most common mutant genes TP53 and CTNNB1 have an effect on diverse biological processes, including metabolism, DNA replication, signaling pathways and cell cycle. The clusters on common enriched pathways explained the function redundancy from putative driver genes with similar function. Our research association with expression profiles investigated the downstream effect of putative driver genes and provided insights about the mechanism for HCC tumor development.

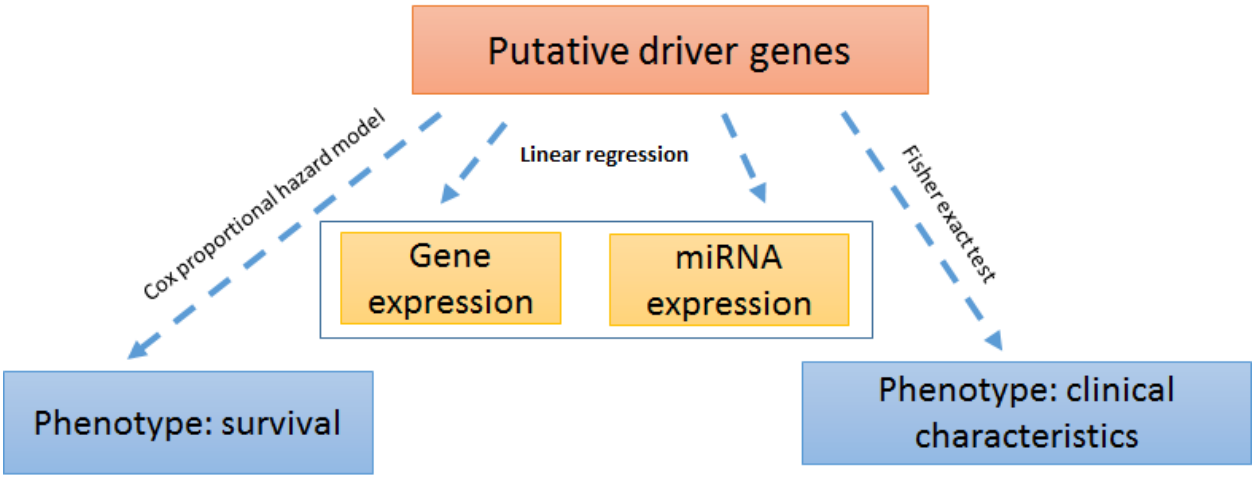

Figure 9: Workflow for the putative driver gene analysis

3.1 Detection of putative driver genes

Using 198 paired tumor-adjacent tissue HCC exome-seq data from TCGA, we obtained 13 genes with "mutation significance" per MutsigCV (2CV v3.1)(Lawrence et al. 2013) (Table 6). Mutation significance can be conceived as an improved metric of mutation frequency, by estimating the personalized and gene-specific background mutation rate. Over all, these putative driver genes are prevalent in 78.3% of the population (Figure 10A). TP53 and CTNNB1 are most significantly mutated genes based on either mutation frequencies or significance (Figure 10A and C), consistent with the observations from other cohorts (Ahn et al. 2014; Shibata and Aburatani 2014). On the other hand, some low-frequency mutation genes have gained rankings in significance per MutsigCV (Figure 10C). For examples, AXIN and BAP1 have 4.55% and 5.56% mutation frequencies, but are ranked 4th and 5th (p-values of 2.22e-04 and 3.19e-04) for their significance. We also detected some interesting putative mutation genes that were not previously well studied and have lower-frequencies (less than 5%), such as EEF1A1, IL6ST and KIF19. The mutual exclusivity among the majority of the driver genes is clearly evident (Figure 10B), especially among the top mutated genes. Interestingly, we detected that CTNNB1 and

ALB mutations frequently co-occur, demonstrated by significant Fisher's exact test (p-value=0.00815) and mean odds ratio 4.2 (95% confidence interval: 1.61-14.56) in the HCC patients. Beyond gene pairs, we detected exclusivity further on gene sets using Dendrix (Vandin, Upfal, and Raphael 2012), and found that up to five top genes (TP53, CTNNB1, BAP1, RB1, AXIN1) have significant mutation exclusivity (p-value < 1e-16).

Table 6: Significantly Mutated Genes in 198 HCC samples

Gene	Description	No. of patients	No. of sites	Mutation Significance (q-value)
TP53	tumor protein p53	62	50	9.13E-13
CTNNB1	catenin (cadherin-associated protein), beta 1, 88kDa	51	26	9.13E-13
RB1	retinoblastoma 1 (including osteosarcoma)	15	17	2.76E-10
AXIN1	axin 1	9	10	0.000222
BAP1	BRCA1 associated protein-1 (ubiquitin carboxy-terminal hydrolase)	10	10	0.000319
TSC2	tuberous sclerosis 2	9	9	0.000574
ARID1A	AT rich interactive domain 1A (SWI-like)	16	16	0.001583
IL6ST	interleukin 6 signal transducer (gp130, oncostatin M receptor)	7	7	0.002616
ALB	albumin	18	20	0.002814
HNF1A	HNF1 homeobox A	8	12	0.018688
APOB	apolipoprotein B (including Ag(x) antigen)	24	26	0.018688
EEF1A1	eukaryotic translation elongation factor 1 alpha 1	5	3	0.071812
KIF19	kinesin family member 19	10	10	0.089399

Functionally, these genes are involved in a wide variety of biological processes, based on their memberships in 25 KEGG pathways (Figure 10D). CTNNB1 and TP53 are involved in many more (10 and 9) pathways as compared to the other genes, confirming their functional importance. Among all 25 KEGG pathways, signaling pathways regulating pluripotency of stem cells have the most number (4) of driver genes, including CTNNB1, HNF1A (Hepatic Nuclear Factor 1 Homeobox A), AXIN1 (Axis Inhibition Protein 1) and IL6ST (Interleukin 6 Signal Transducer).

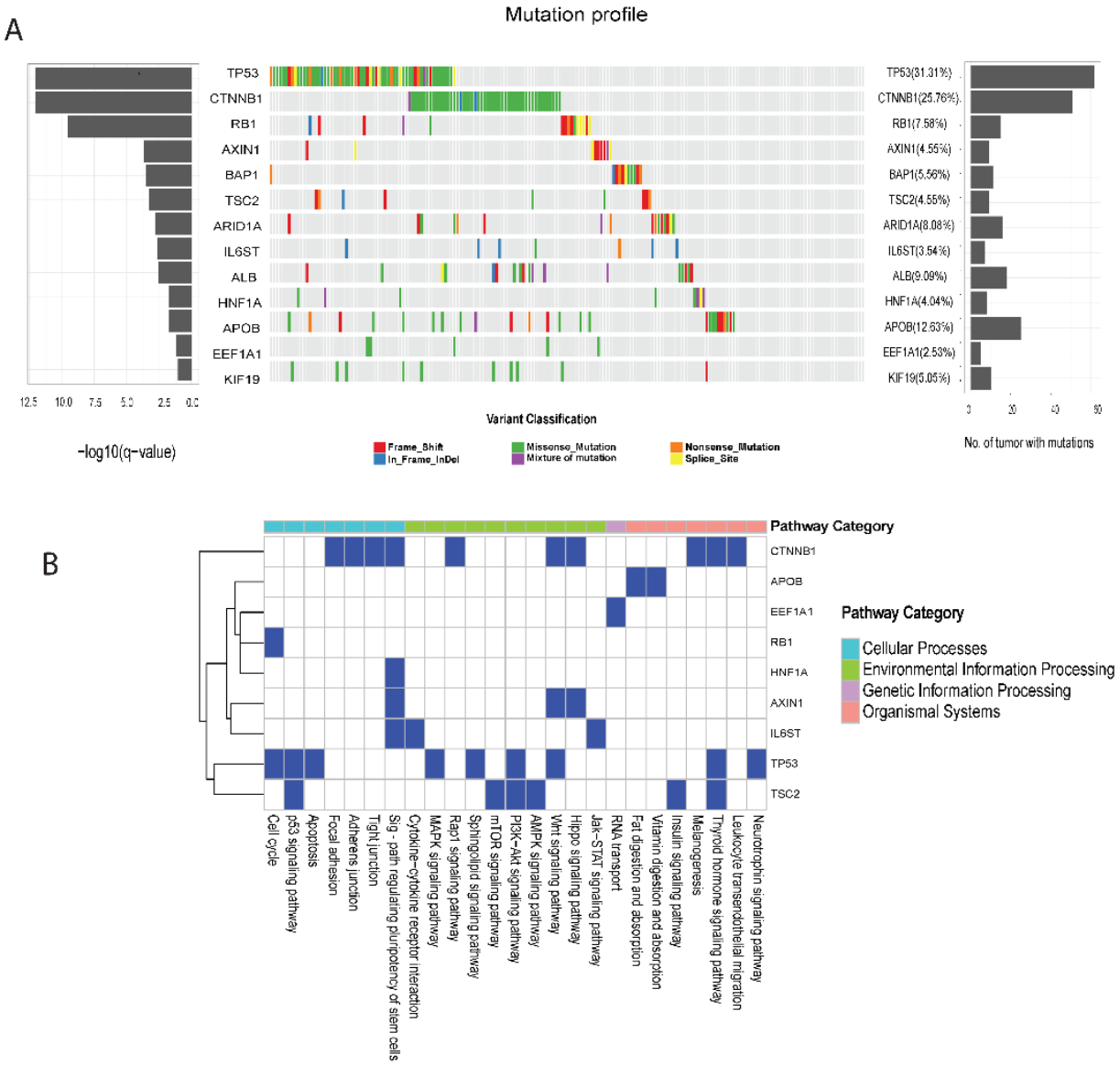

Figure 10: Putative driver genes and involved KEGG pathways

(A) Mutation profiles of these putative driver genes. The left part indicating mutation significance, the middle part indicating variant classifications and the right part indicating mutation frequency. (B)The putative driver gene involved KEGG pathways.

To investigate the landscape of driver mutations at the protein level, we annotated the variants in the putative driver genes and present their distributions in lollipop plots (Figure 11). The point mutations are very sparse among HCC patients with variable positions. However, a few relatively hot spots in TP53, CTNNB1 and EEF1A1 do emerge. The highest (12) point mutation occurs in amino acid position D32G/N/V/Y in CTNNB1. It was reported to relate to a high level of ß-catenin activation during HCC progression (Rebouissou et al. 2016). R249S mutation in TP53 occurs in 8 samples, and it was reported frequently mutated in HCC patients exposed to aflatoxin B1 (AFB1) (Aguilar, Hussain, and Cerutti 1993). This mutation induces substantial structural perturbation around the mutation site in the L3 loop of TP53 (Joerger et al. 2005; Friedler et al. 2004). Interestingly, all mutations in eukaryotic elongation factor 1A1 (EEF1A1) occurs at one position T432I/L/S. The T->S mutation in HCC was recently reported by others (Schulze et al. 2015). The amino acid substitutions are possibly related to the lower expression of EEF1A1 in mutant samples. We also observed that the majority of mutations are detrimental truncating mutations, rather than missense or in-frame mutations. A good example is BAP1, where 7 out 11 mutation types are truncating mutations.

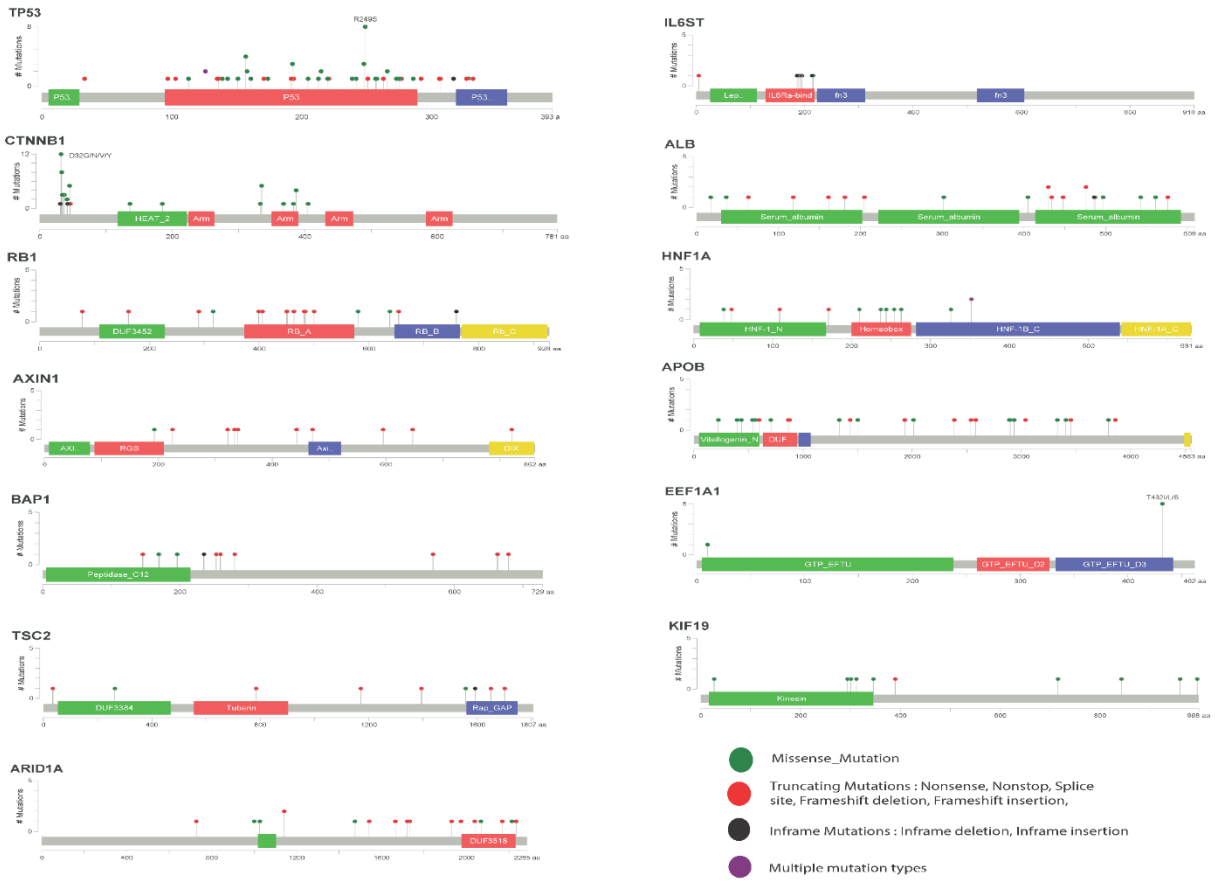

Figure 11: Mutation effect at the protein level

3.2 Associations among clinical characteristics and putative driver genes

To reveal the possible associations of these driver genes with physiological and clinical characteristics of patients, such as risk factors, gender, race and grade, we conducted Fisher's exact tests among them (Figure 12). Regarding the associations with risk factors, CTNNB1 mutation is enriched (39.2%, p=0.0492) in alcoholic patients, similar to the previous report (Guichard et al. 2012), while RB1 mutations are related to hepatitis B infection. For gender associations, TP53 and CTNNB1 are more frequently mutated in males than females. As for racial preference, TP53 and RB1 are associated with Asians while ALB mutations are more

prevalent in Caucasians. Additionally, TP53 mutation frequency tends to increase with higher tumor grades. Our findings show that complex disparities exist among driver genes.

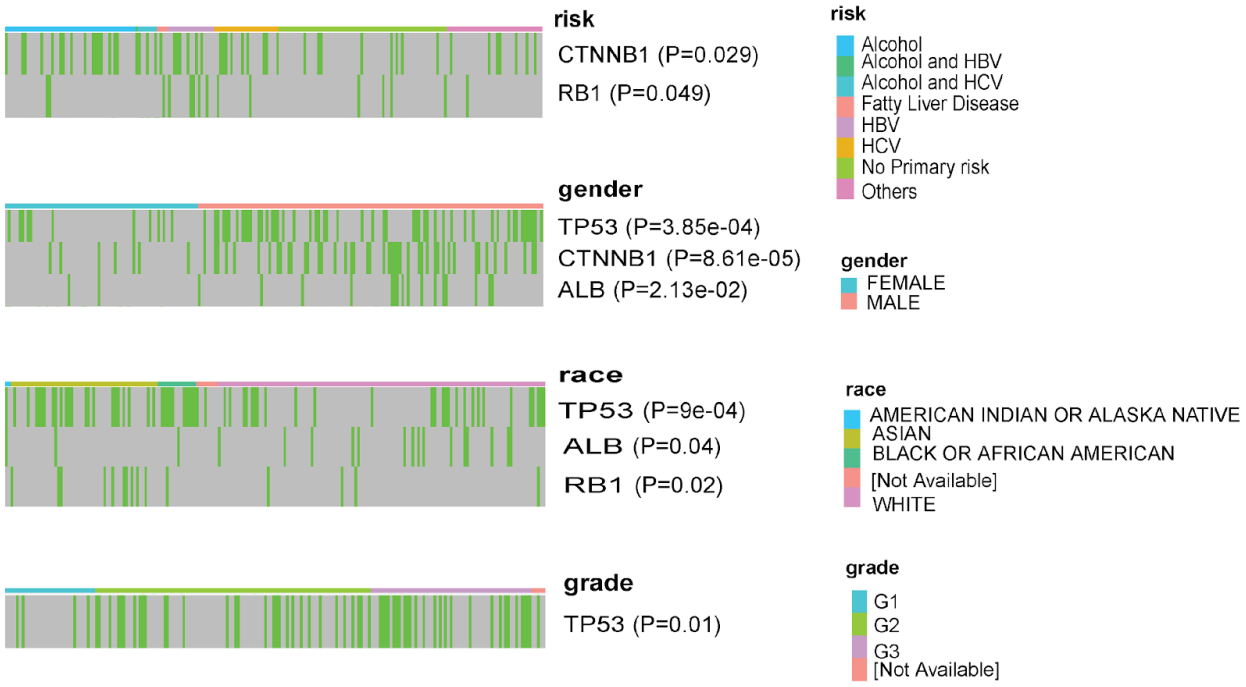

Figure 12: Putative driver genes associated with clinical characteristics

Shown are genes with significant associations to clinical features, including risk, gender, race and grades (Fisher's exact p-value< 0.05). Patients with mutation of driver genes of interest are marked by green color.

In order to test survival associations from all the driver mutations, we built a multivariable Cox-PH model on the relapse-free survival. We used the median of fitted prognosis index score generated from Cox-PH model and divided samples into high and low risk groups. The survival difference between the two risk group is significant (p-value=0.01) with decent concordance index (CI=0.616), indicating the mutations of the 13 putative genes are reliable features for the survival model (Figure 13). Meanwhile, we obtained survival difference on overall survival with CI 0.667 and p value 0.027.

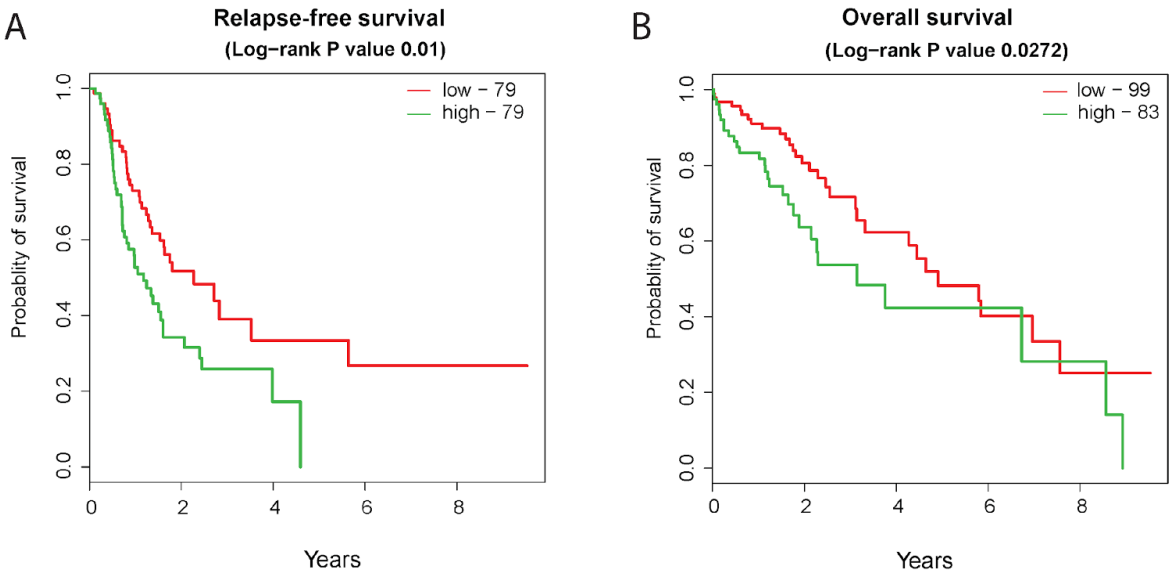

Figure 13: Multiple variable survival model on putative driver genes

The cox regression from R *glmnet* package was used to build the survival model featuring the putative driver genes by ridge regression. The samples were divided into high and low risk groups by median prognosis index. Kaplan-Meier survival curves were plotted for these two groups on relapse-free survival (A) and overall survival (B).

Although we did not expect any single driver gene dominantly contribute to prognosis outcome in the population, we did observe that hazard ratios (HR) are high in some putative driver genes, including APOB and EEF1A1 (Fig S3). Notably, the mutation of EEF1A is associated with higher hazard ratio of 4.621 (log-rank p-value=0.005). Intriguingly, the IL6ST mutations are associated with good prognosis compared to the wild group (log-rank p-value=0.007). We observed the similar hazard ratio trends for individual driver genes in the relapse-free survival, though not statistically significant. All together, the analysis show that the driver genes collectively may differentiate HCC survival outcome. (Figure 14)

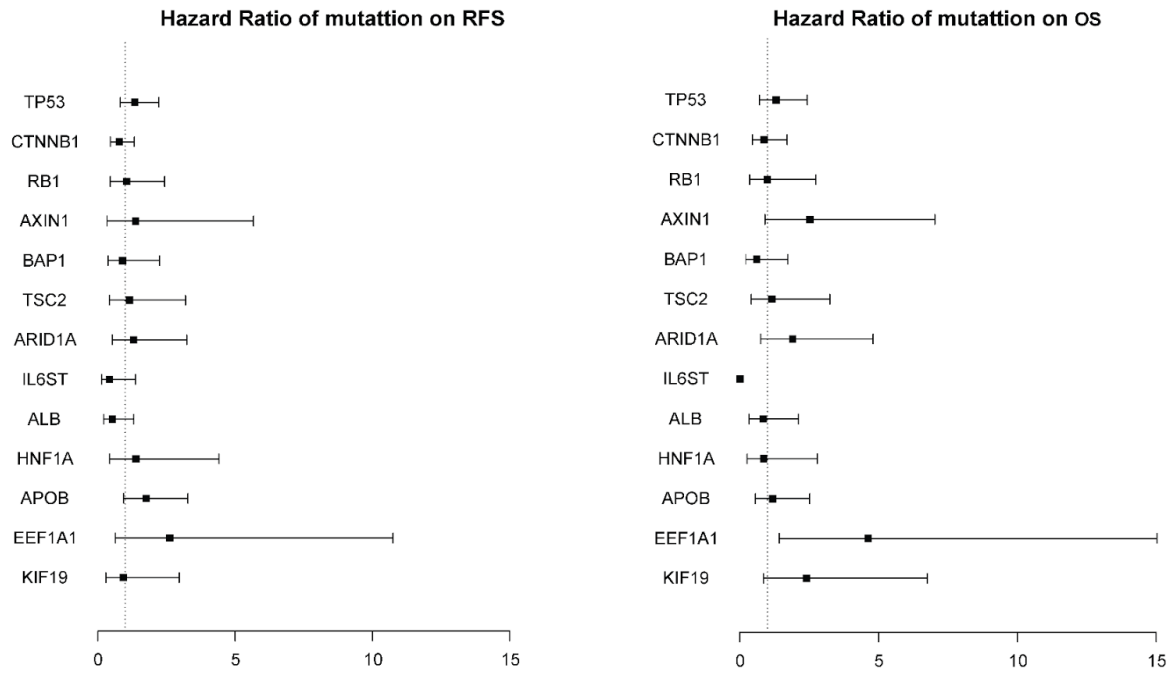

Figure 14: Hazard Ratios of the putative driver gene mutation

3.3 The associations between mRNA transcriptome and putative driver gene mutations

We first examined the cis-effect of these driver mutations on their own gene expression levels. Most driver gene mutations in HCC samples have decreased expression levels of these genes, except CTNNB1 which has increased expression values ($p=0.005$). To obtain the lists of genes impacted by the putative drivers at the transcriptional level, we built linear models using mutation statuses of putative driver genes as the predictors and individual gene expression values as responses, similar to others' studies (Gerstung et al. 2015). Doing so, the potential confounding effects among putative driver genes are minimized by the model. Moreover, the putatively affected genes were verified to be statistically significant based on the background distribution, generated by random sampling of mutated 13 driver genes. Our results reveal that over 40% (9130) of genes are significantly associated (FDR<0.05) with putative driver genes. We also tabulates the number of genes significantly associated to each putative driver gene

(Figure 15A). The top two mutated genes TP53 and CTNNB1 both affect over two thousand genes. Additionally, although BAP1 has a low mutation rate of 5.56%, it is ranked 3rd and linked to gene expression changes in over 1700 genes.

To investigate the biological processes these 9130 genes are involved in, we conducted KEGG pathway enrichment analysis and detected 73 significantly impacted pathways with BH adjusted p-values <0.05 (Figure 15B). We further categorized these pathways into 5 super groups according to the KEGG pathway organization, namely: cellular processes, environmental information processing, genetic information processing, metabolism, and organismal systems. Among the driver genes, CTNNB1, RB1, TP53 and BAP1 are most densely connected to enriched pathways from associated genes. BAP1 affects far more metabolism related functions that were overlooked previously. The network of putative driver genes and affected pathways present explanations for mutual exclusivities observed earlier, in that mutual exclusive genes share similar pathways. Among TP53, CTNNB1, BAP1 and RB1, TP53-RB1 are both involved in cell cycle pathway and DNA replication; BAP1-TP53 both affect many amino-acid metabolic and lipid synthesis pathways; and BAP1-CTNNB1 share processes such as protein digestion and absorption. Besides, CTNNB1-HNF1A are both involved in drug, steroid and retinoid metabolism.

It is not surprising that the pathway super-group affected most by the putative driver genes are metabolic pathways. However, we also observe that some signaling pathways in the environmental information processing group are significantly influenced by driver genes CTNNB1, BAP1 and RB1. ECM-receptor interaction and Wnt pathways are associated with CTNNB1 mutations; Hedgehog signaling pathway and ABC transporter pathway are associated with BAP1 mutation; and TGF-beta pathway is associated with both CTNNB1 and RB1 mutations. On the other hand, some previously less studied driver genes in HCC, such as TCS2, KIF19 and ARID1A, are shown to associate with sugar-group modification and metabolism, cell

cycle and metabolic pathways.

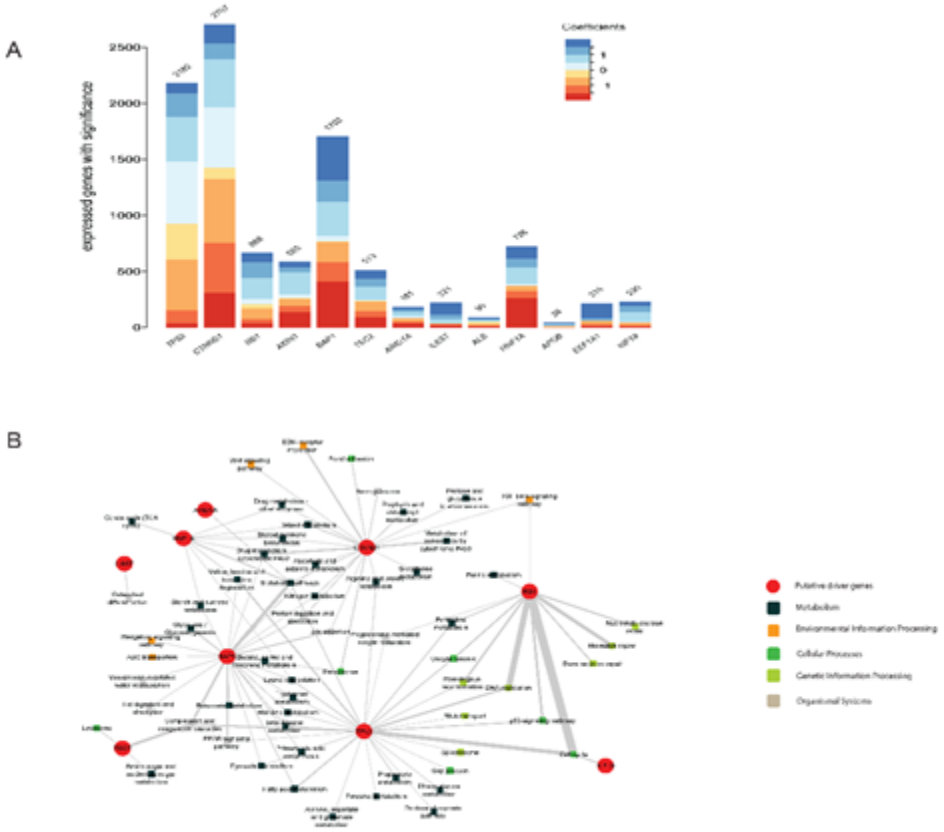

Figure 15: Associations of putative driver genes with gene expression

(A) The number of genes whose expression values are significantly associated with the driver genes, divided by mutation frequency. (B) enriched KEGG pathways among significant genes as shown in (A).

3.4 The associations between miRNA expression and putative driver gene mutation

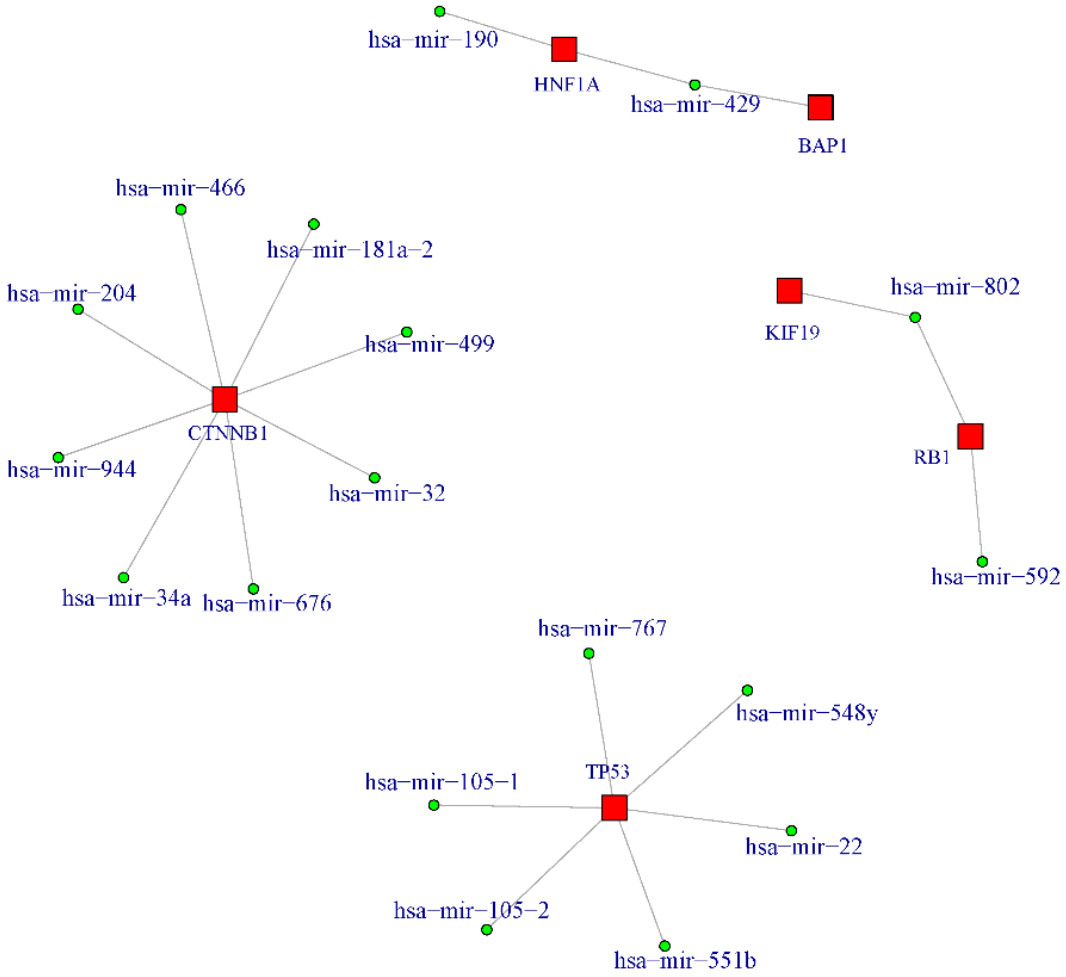

Figure 16: The network of impacted miRNAs and driver genes

To extend the investigation of putative driver genes' effect to miRNAs, we built the linear models between driver genes and miRNA expression values, similar to mRNA expression analysis above. The resulting network analysis shows that six putative driver genes have significant effects on 18 miRNAs (Figure 16). The two most significantly mutated genes TP53 and CTNNB1 have the most numbers of affected miRNAs. TP53 is associated with 6 miRNAs: miR-1051 and -2-1/2, miR-551b, miR-22, miR-548y, and miR-767. CTNNB1 is associated with changes in 8 miRNAs: miR−466, miR−181a−2, miR−204, miR−944, miR−34a, miR−676, miR−32 and miR−499. Other four driver genes are associated with expression levels of four other miRNAs. HNF1A mutation is associated with expression levels of miR-190 and miR-429, the latter of which is linked to BAP1 mutation too. RB1 is associated with miR-592 and miR-809, the latter of which is linked to KIF19. The cancer relevance of several of these miRNAs are supported by literatures. MiR-181a-2, miR-204, miR-32 and miR-429, play roles in various cancers according to miRCancer database (Xie et al. 2013). Among them, miR-429 was reported down-regulated in HCC (You et al. 2013), and may target BAP1 3' UTR with strong evidence from reporter assays (Hyun et al. 2009). MiR-22 was reported up-regulated in HCC (Jiang et al. 2011), and it may target TP53 3' UTR. Specifically, miR-32 was reported to target PTEN to promote proliferation in HCC (Yan et al. 2015), while miR-181a-2 was critical in hepatic cancer stem cells (Ji et al. 2009). Our results indicate that mir-32 and mir-181a-2 dysregulation may arise from CTNNB1 mutations.

Table 7: Functional annotation on the miRNAs

miRNA	Cancer	Profile	PubMed.Article
hsa-mir-181a-2	glioma	down	MicroRNA-181 inhibits glioma cell?proliferation by targeting cyclin B1.
hsa-mir-181a-2	hepatocellular carcinoma	up	Identification of microRNA-181 by genome-wide screening as a critical player in EpCAM-positive hepatic cancer stem cells.
hsa-mir-181a-2	non-small cell lung cancer	down	MicroRNA-181 functions as a tumor suppressor in non-small cell lung cancer (NSCLC) by targeting Bcl-2.
hsa-mir-181a-2	papillary thyroid carcinoma	up	Expression of miRNAs in Papillary Thyroid Carcinomas Is Associated with BRAF Mutation and Clinicopathological Features in Chinese Patients.

hsa-mir-181a-2	prostate cancer	up	microRNA-181 promotes prostate cancer cell proliferation by regulating DAX-1 expression.
hsa-mir-204	breast cancer	down	MicroRNA-204 targets JAK2 in breast cancer and induces cell apoptosis through the STAT3/BCl-2/survivin pathway.
hsa-mir-204	gastric cancer	down	MiR-204 down regulates SIRT1 and reverts SIRT1-induced epithelial-mesenchymal transition, anoikis resistance and invasion in gastric cancer cells.
hsa-mir-204	glioma	down	Loss of miR-204 expression enhances glioma migration and stem cell-like phenotype.
hsa-mir-204	intrahepatic cholangiocarcinoma	down	miR-204 inhibits epithelial to mesenchymal transition by targeting slug in intrahepatic cholangiocarcinoma cells.
hsa-mir-204	malignant melanoma	down	Regulation of cancer aggressive features in melanoma cells by microRNAs.
hsa-mir-204	nasopharyngeal carcinoma	down	Down-regulation of miRNA-204 by LMP-1 enhances CDC42 activity and facilitates invasion of EBV-associated nasopharyngeal carcinoma cells.
hsa-mir-204	non-small cell lung cancer	down	miR-204 functions as a tumor suppressor by regulating SIX1 in NSCLC.
hsa-mir-204	non-small cell lung cancer	down	MiR-204 inhibits human NSCLC metastasis through suppression of NUAK1.
hsa-mir-204	osteosarcoma	down	MicroRNA-204 inhibits proliferation, migration, invasion and epithelial-mesenchymal transition in osteosarcoma cells via targeting Sirtuin 1.
hsa-mir-204	prostate cancer	up	Mechanisms and functional consequences of PDEF protein expression loss during prostate cancer progression.
hsa-mir-204	renal cell carcinoma	down	Upregulation of microRNA-204 inhibits cell proliferation, migration and invasion in human renal cell carcinoma cells by downregulating SOX4.
hsa-mir-204	renal clear cell carcinoma	down	VHL-regulated MiR-204 suppresses tumor growth through inhibition of LC3B-mediated autophagy in renal clear cell carcinoma.
hsa-mir-204	retinoblastoma	down	MiR-204, down-regulated in retinoblastoma, regulates proliferation and invasion of human retinoblastoma cells by targeting CyclinD2 and MMP-9.
hsa-mir-32	acute myeloid leukemia	up	MicroRNA-32 upregulation by 1,25-dihydroxyvitamin D3 in human myeloid leukemia cells leads to Bim targeting and inhibition of AraC-induced apoptosis.
hsa-mir-	colorectal cancer	up	The relationship between and clinical significance of

32			MicroRNA-32 and phosphatase and tensin homologue expression in colorectal cancer.
hsa-mir-32	hepatocellular carcinoma	up	MiR-32 induces cell proliferation, migration, and invasion in hepatocellular carcinoma by targeting PTEN.
hsa-mir-32	non-small cell lung cancer	down	Expression of miR-32 in human non-small cell lung cancer and its correlation with tumor progression and patient survival.
hsa-mir-32	non-small cell lung cancer	down	Tanshinones suppress AURKA through up-regulation of miR-32 expression in non-small cell lung cancer.
hsa-mir-32	non-small cell lung cancer	down	miR-32 functions as a tumor suppressor and directly targets SOX9 in human non-small cell lung cancer.
hsa-mir-32	oral squamous cell carcinoma	down	MiR-32 functions as a tumor suppressor and directly targets EZH2 in human oral squamous cell carcinoma.
hsa-mir-32	osteosarcoma	down	MicroRNA-32 inhibits osteosarcoma cell proliferation and invasion by targeting Sox9.
hsa-mir-429	non-small cell lung cancer	down	Expression of miR-29c, miR-93, and miR-429 as Potential Biomarkers for Detection of Early Stage Non-Small Lung Cancer.
hsa-mir-429	oral squamous cell carcinoma	down	MiR-429 inhibits oral squamous cell carcinoma growth by targeting ZEB1.
hsa-mir-429	osteosarcoma	down	Tumor-Suppressing Effects of miR-429 on Human Osteosarcoma.
hsa-mir-429	ovarian cancer	down	A miR-200 microRNA cluster as prognostic marker in advanced ovarian cancer.
hsa-mir-429	ovarian cancer	down	Ectopic over-expression of miR-429 induces mesenchymal-to-epithelial transition (MET) and increased drug sensitivity in metastasizing ovarian cancer cells.
hsa-mir-429	prostate cancer	up	Downregulation of microRNA-429 inhibits cell proliferation by targeting p27Kip1 in human prostate cancer cells.
hsa-mir-429	renal cell carcinoma	down	MicroRNA-429 suppresses cell proliferation, epithelial-mesenchymal transition, and metastasis by direct targeting of BMI1 and E2F3 in renal cell carcinoma.

Chapter 4. Multi-omic data integration to stratify population in hepatocellular carcinoma

Building robust subgroups helps to guide precise targeted therapeutics. Integrating different levels of omic datasets makes it possible to stratify patients and discover distinct features for each subgroup. However, such comprehensive integration of multi-omic data has been lacking in HCC studies. With TCGA multi-omic data, we performed integrative clustering analysis of 360 HCC samples downloaded from TCGA, using the information of DNA copy number variation, CpG methylation, mRNA and miRNA expression. We discovered five molecular subtypes with significant differences in terms of survival. These subtypes function as an independent predictive feature on patient survival, apart from clinical characteristics. Furthermore, we confirmed the multi-omics predictive model on each omic level, and the concordance between them ranges 56-87%. This shows that that multi-omics classification results are applicable on new samples, even when they have fewer than four omic datasets. Additionally, we identified gene expression signature and active biological pathways of these subtypes. The analysis was done in the flow (Figure 17).

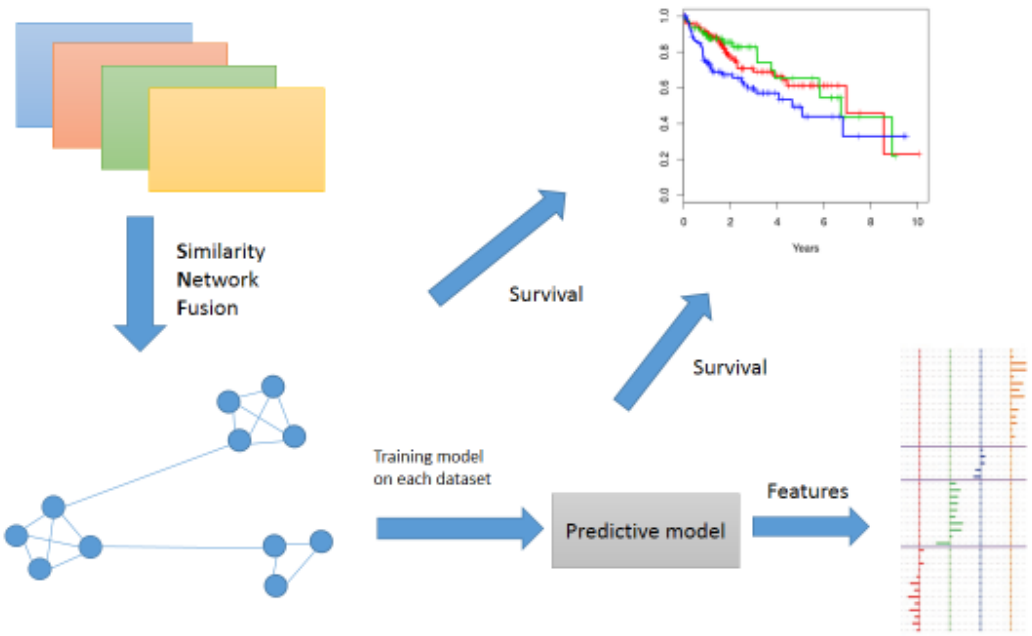

Figure 17: Workflow for the integrative analysis

4.1 Integrative clustering of HCC multi-omics data

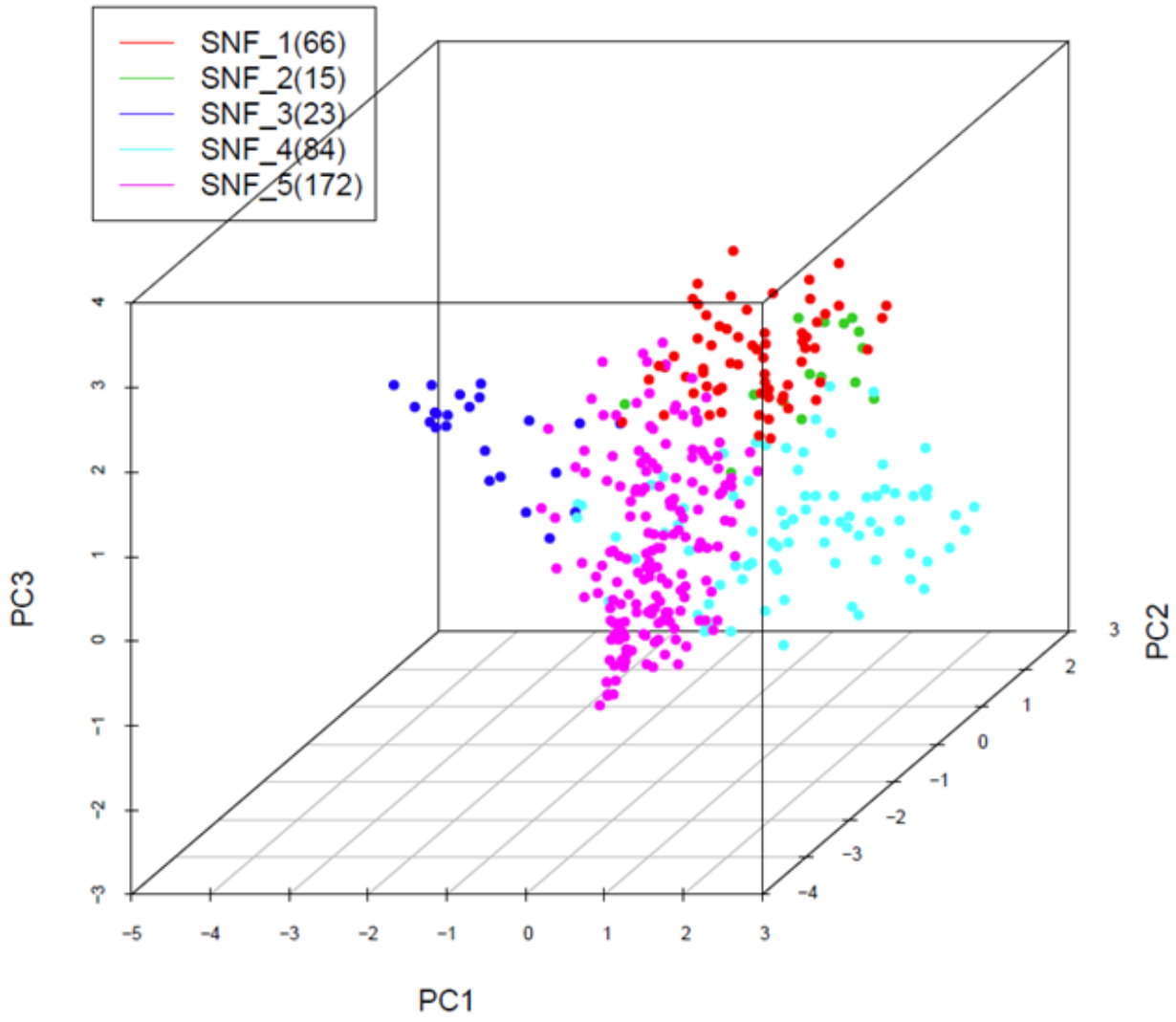

Figure 18: HCC subtypes from multi-omic datasets

From the TCGA liver HCC project, we obtained 360 tumor samples, along with 39 tumor adjacent normal samples, which have RNA-seq, miRNA-seq, DNA methylation and copy number alterations. For these 399 samples, we did pre-processing as described in the Materials and Methods section, and obtained 20167 (of 20531) genes from RNA-seq, 854 (of 1046)

miRNAs from miRNA-seq, 20150 (of 20772) genes from DNA methylation and 11550 (of 24958) genes with copy number alterations.

We assume that different levels of omics information contain complementary information important to investigate the mechanism of cancer initiation and development. We used Similarity Network Fusion (SNF) (B. Wang et al. 2014), a newly developed computational approach for data integration of different omic data sets. This method generate a sample similarity matrix on each omics layer of the HCC samples, and then integrates sample similarity matrices from different omic layer iteratively into a comprehensive sample similarity matrix using graph fusions. This approach allows to minimize feature selection bias, noise and barriers from different data types.

Before building the subclusters in HCC samples, we first performed proof-of-concept test to see the integrative similarity matrix could contain the representative information to distinguish tumor samples and tumor adjacent normal samples (Figure 19). We used one unsupervised spectral clustering approach, and assigned these samples into 2 subgroups. Similarly, we built 2 subgroups on each of these four omic datasets. As results, the 39 tumor adjacent normal samples are correctly assigned into one group in the three datasets of gene expression, DNA methylation and copy number variances, based on the similarity. In the miRNA expression data layer, a small fraction (5 out of 39) of normal samples are assigned incorrectly with the rest of tumor group. However, in the fused similarity matrix, we observe a clearer distinguishing between normal samples and tumor samples. This result shows that indeed the multi-omics fusion is efficient to integrate different levels of omic information. Some HCC samples have closeness to normal samples, which might suggest better outcome in clinic.

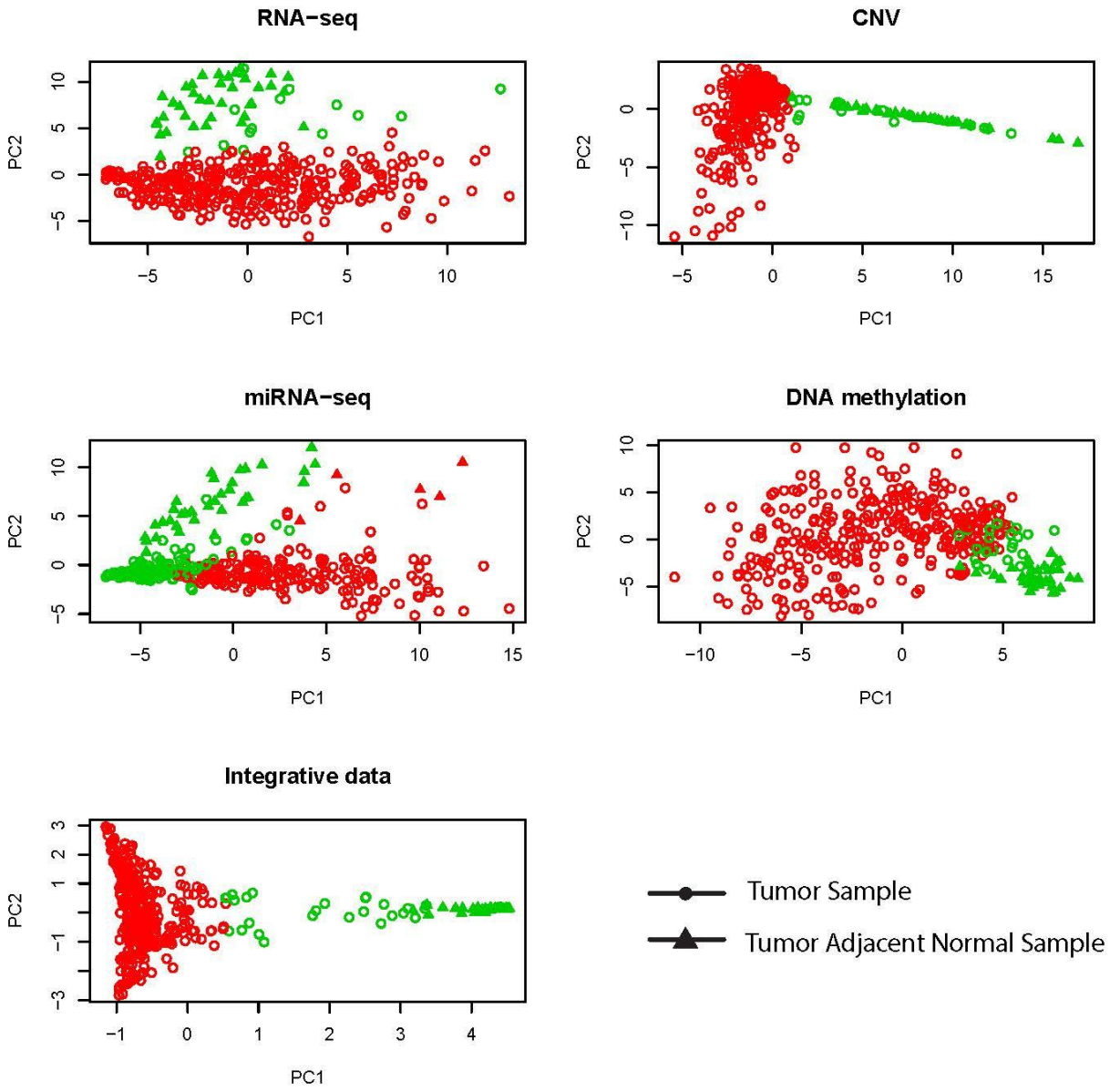

Figure 19: unsupervised clustering of tumor tissues and tumor-adjacent tissues on affinity matrix

We next conducted the essential sub-clustering identifications on 360 HCC samples, along with 39 tumor adjacent samples. To explore optimal numbers of sub-groups, we tested the cluster number from k=2 to 8 through spectral clustering and evaluated the accuracy using Dunn's test, PCA and Adjusted Rand Index (ARI) (Figure 20 and 21). Dunn's test shows that all the sub-

clusters have values over 0.9 (Figure 20). Additionally, 3D Principal Component Analysis (PCA) plot confirmed that segregation of different groups are sensible (Figure 21). On the other hand, ARI shows that different number of clusters are well-related, but not replaceable, except that cluster number k=2 has low correlations with other clusters (Figure 20).

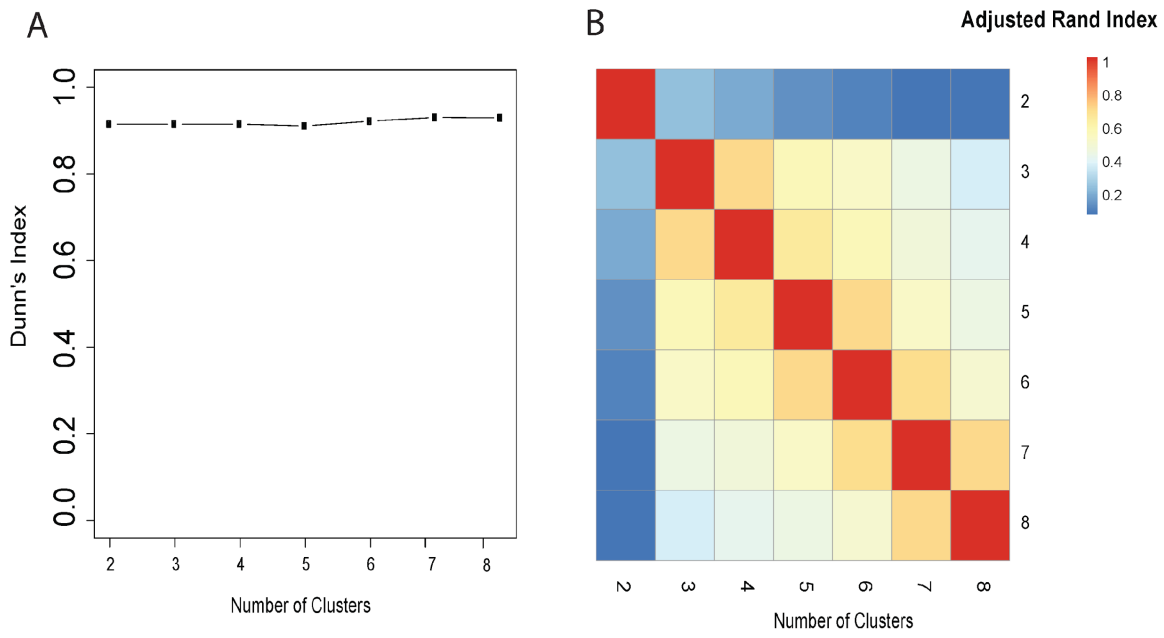

Figure 20: statistics on unsupervised clustering on different number of clusters. (A) Dunn's Index, (B) The adjusted Rand index (ARI)

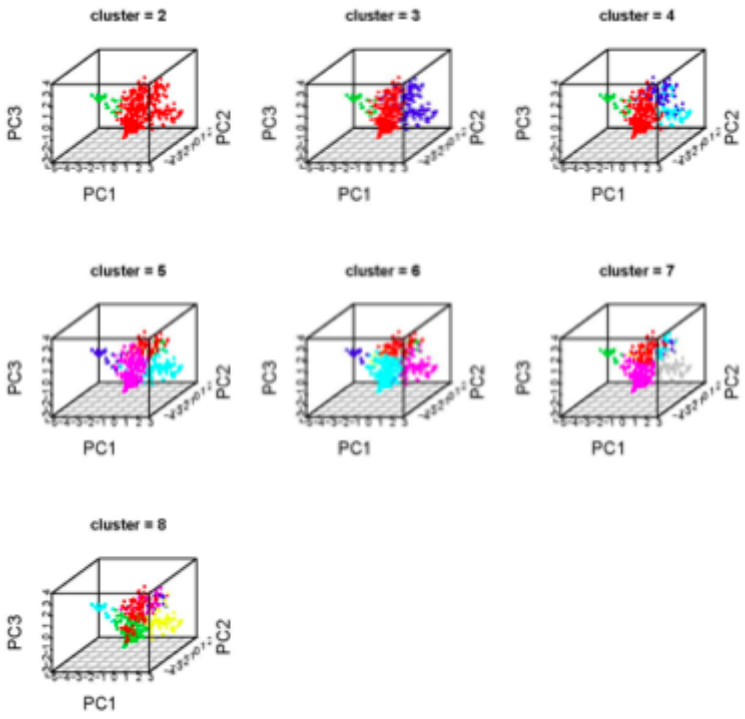

Figure 21: PCA 3D on different number of clusters

In the light of similar accuracies on sub-clusters k=2 to 8 based genomics information, we sought additional guide from survival information to determine the optimal k. We built Cox-PH models of overall survival (OS) and relapse-free survival (RFS) fitted by the sub-clusters k, and used log rank tests to indicate the significance of each model (Figure 22). We optimized k=5 subtypes on these HCC samples, since OS (P=0.03) and RFS (P=0.01) Cox-PH models give the best p-values combined. As a quality control, we observe that tumor adjacent normal samples are assigned into one group, SNF_3, indicating that sub-cluster 3 is expected to have good survival. We visualize the 3D PCA plot of 5 subgroups in Figure 18, and name subtypes as SNF_1 to SNF_5.

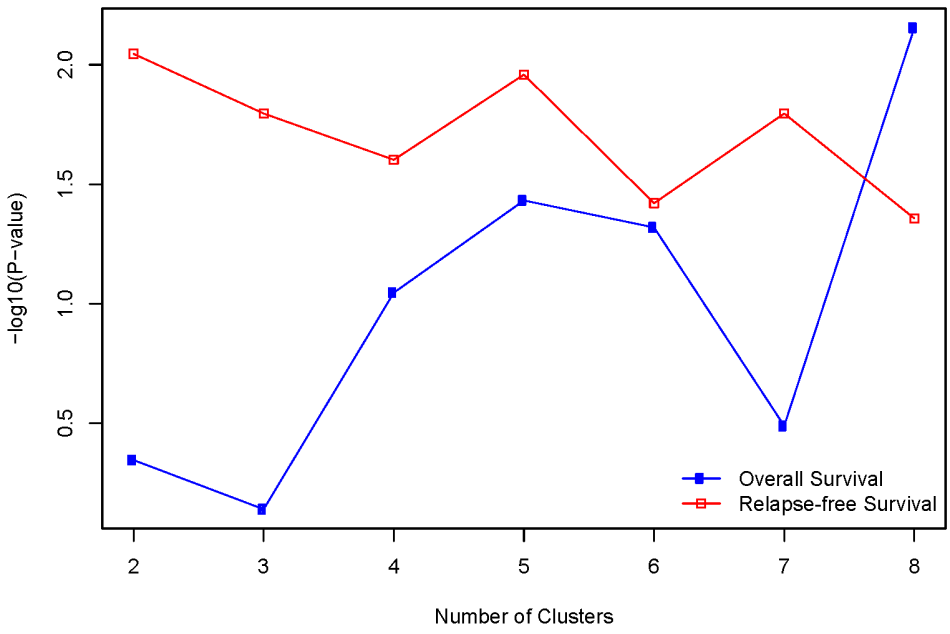

Figure 22: Survival Association of different number of clusters

4.2 The associations of SNF sub-clusters with clinical characteristics

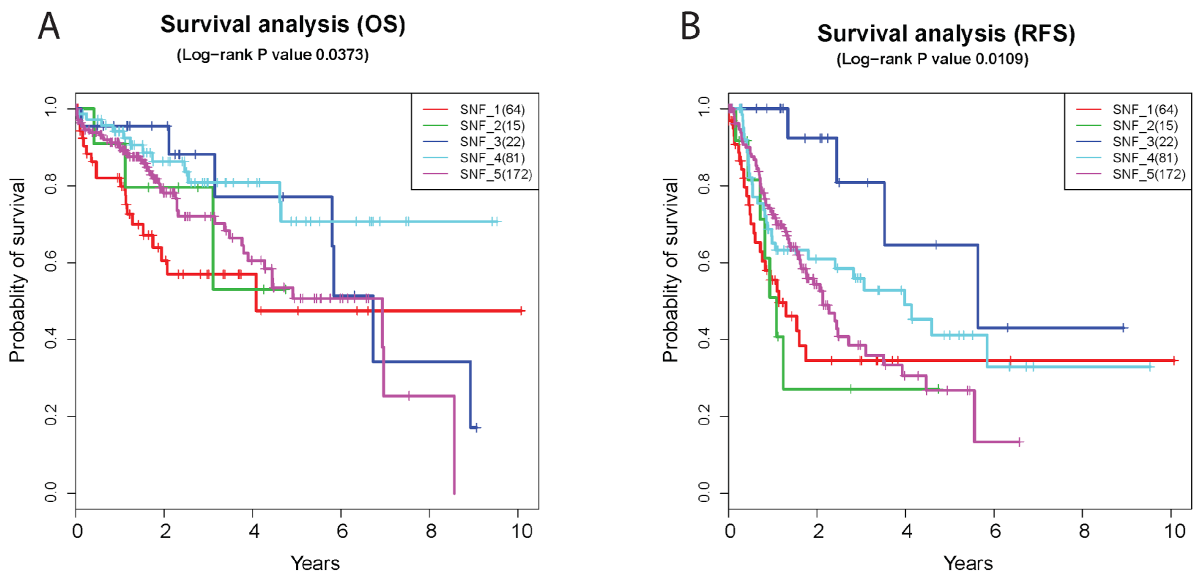

Figure 23: HCC subtypes associated with Survival

(A) Kaplan-Meier estimate curves of overall survival among patients in 5 subtypes. (B) Kaplan-Meier estimate curves of relapse-free survival among patients in 5 subtypes.

The fused sample similarity matrix are based on genomics information and does not have input from clinical characteristics. Rather, we used survival information *post hoc* to determine the best sub-cluster number k=5. The OS Cox-PH model with k=5 gave a Log-rank p value 0.03 with Concordance Index 0.605 (measurement of correlations among sub-clusters from the data, as compared to the random background). Among the 5 sub-types, SNF_1 has the worst OS with an average of 14.8 months, while SNF_3 has the best survival of an average of 27.9 months. Similarly, in RFS analysis SNF_1 has the shortest average 10 months to relapse while SNF_3 has a poor average of 25.7 months to relapse.

We also tested the associations of our identified subtypes with other clinical characters, including gender, race, grade, stage and risk. The result showed that our identified subtypes are significantly associated with gender, race, grade, stage and risks (Table 8). In addition, we found that SNF_3 group has the majority of early stage and low grade HCC cases, which would help to explain have their closest similarity to the normal adjacent samples and better survival.

Table 8: Subtypes distribution in clinical characteristics

Clinical Characteristics	Varitables	SNF_1	SNF_2	SNF_3	SNF_4	SNF_5
Gender (p = 0.00001)	FEMALE	22	11	9	38	32
	MALE	42	4	13	43	140
Race (p = 0.01089)	WHITE	31	10	17	32	84
	ASIAN	28	5	3	47	69
	BLACK OR AFRICAN AMERICAN	5	0	2	1	9
	AMERICAN INDIAN OR ALASKA NATIVE	0	0	0	0	1
	[Not Available]	0	0	0	1	9
Grade (p = 0.00055)	G1	5	1	9	4	33
	G2	26	8	9	37	89
	G3	30	4	4	35	44
	G4	3	1	0	4	4
	[Not Available]	0	1	0	1	2
Stage (p = 0.02985)	Stage I	18	3	13	44	88
	Stage II	18	4	6	17	39
	Stage III	23	6	2	15	31
	Stage IV	1	0	0	2	2
	[Not Available]2	4	2	1	3	12
Risk (p = 0.00291)	Alcohol	18	4	6	13	25
	Alcohol and HBV	1	0	0	4	15
	Alcohol and HCV	2	0	0	1	11
	Fatty Liver Disease	0	0	1	2	7
	HBV	9	0	0	26	37
	HCV	6	1	2	3	19
	No Primary risk	17	7	10	20	31
	Others	11	3	3	12	27

To test if the five SNF subtypes has prognosis values in addition to the clinical characteristics, we built a combined molecular and clinical Cox-PH model, and compared with the baseline Cox-PH model on the clinical characteristics (stage, grade, race, gender, age and risk). We used

ANOVA test on the full model (logrank Pvalue 7e-4) and the baseline model (logrank Pvalue 4.3e-5), and we obtained a significant Chi-square P-value 0.0055 on the overall survival while a significant Chi-square P-value 0.034 on the relapse-free survival model. The result shows that our identified subtypes could function as an independent predictor of survival, from other clinical characteristics.

4.3 The associations of SNF subtypes with putative driver genes

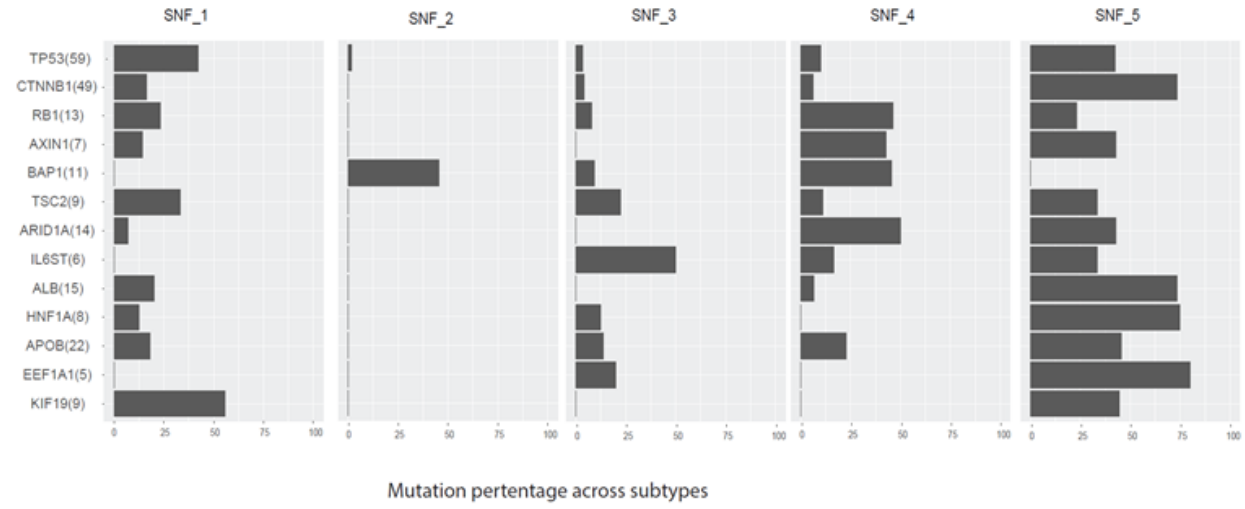

Figure 24: HCC subtypes associated with putative driver genes

We had obtained 13 putative driver genes (Chapter 3), with a fraction of the 360 HCC samples whose exome-seq data are available. Since our SNF subtypes were not trained on exome-seq data but other four omics-data, we next tested if these subtypes have different driver mutation profiles. Indeed, samples from each subtype consists of different combinations of putative driver genes (Figure 24). The top two most frequently mutated genes TP53 and CTNNB1 are frequently mutated in SNF_1 and SNF_5. Almost half of the BAP1 mutant samples are in SNF_2 group, in which the rest of driver genes were almost all wild types. For SNF_3, IL6ST is the major driver gene, which may contribute to the best survival. SNF_5 contains diverse mutation in driver genes, where the majority of CTNNB1, ALB, HNF1A and EEF1A1 are mutant.

4.4 SNF subtype predictive model on each omics dataset

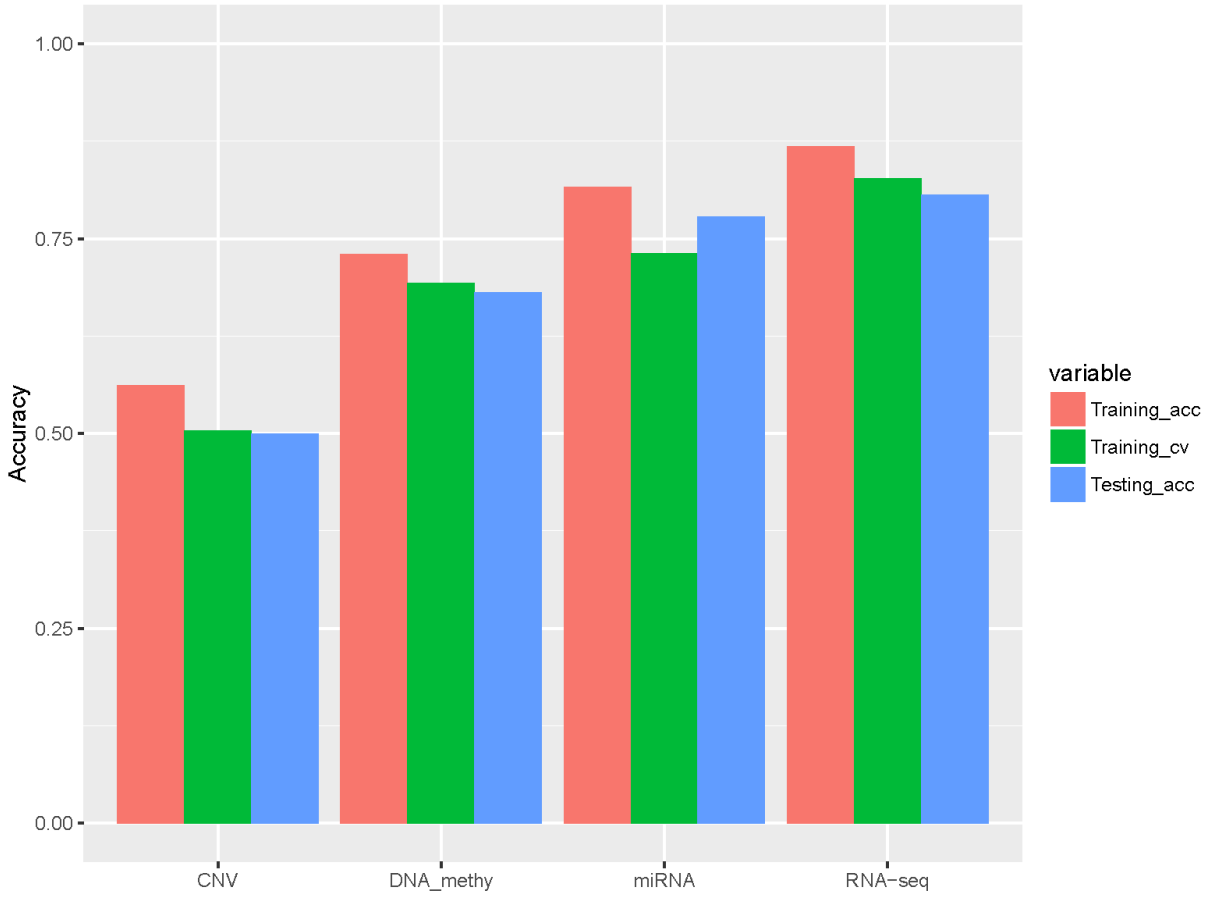

Figure 25: The concordances between each omics data and the multi-omics predictive model

Generally, a large cohort may not have all four omics types given the high cost of NGS. The subtypes identified with multi-omics data should still be applicable to new patients with just one or a few (n<4) omic data types. To demonstrate this competency, we next projected the fusion information into each dataset and build predictive models on that perticular dataset. For each dataset, we divided all samples into the training data (80%) and the testing data (20%), with the identified subtypes as the response. We exploited the nearest shrunken centroid method, which calculates a class 'centroid' for each subclass and assigns samples to the closest subclass, and trained models on each dataset. We calculated the concordances between predicted subtypes and

our identified subtypes on training data, 10 fold cross validation on tarining data and testing data, as seen in Figure 25.

Among the training data of four omics-sets, the predictive model based on gene expression has the best concordance (87%). While mRNA-seq, DNA methylation and copy number variation have sequentially less accuracy, with 82%, 73% and 56% concordances respectively. Similar but less concordance scores exist in cross validation and testing data. This result confirms that the projection onto one data set is applicable, while preserving majority of the accuracies from the features extracted from the comprehensive multiple omic subtypes.

Additionally, we applied our model on another independent DNA methylation dataset composed of 27 HCC samples and 20 normal adjacent samples (Song et al. 2013). The 20 normal samples were used as quality controls to check if we can successfully cluster them as part of SNF_3, since this group resemble the normal samples the best. As a result, these entire 47 samples were categorized into four subtypes, SNF_1, SNF_3, SNF_4 and SNF_5. These 4 subgroups of new samples show different survival outcomes. SNF_3 cluster were least aggressive with longer survival in the training model. And as expected, 17 out of 20 normal methylation samples are correctly assigned to SNF_3, testifying the accuracy of our multi-omics model. Meanwhile, we used Cox-PH to build a survival model between predicted subtypes (predictors) and overall survival (responses) of these 47 samples, and obtained significant association results (Concordance Index 0.656 and Log-rank P-value 0.009). Along with other clinical characteristics stage Sex, Age, HBV infection and HCV infection, the full model including SNF subtypes have an increasingly significant prediction on survival. This analysis on the independent DNA methylation dataset validates the effectiveness of our classification and groups patients into different outcome groups from the molecular level.

4.5 SNF subtype feature on gene expression profiling

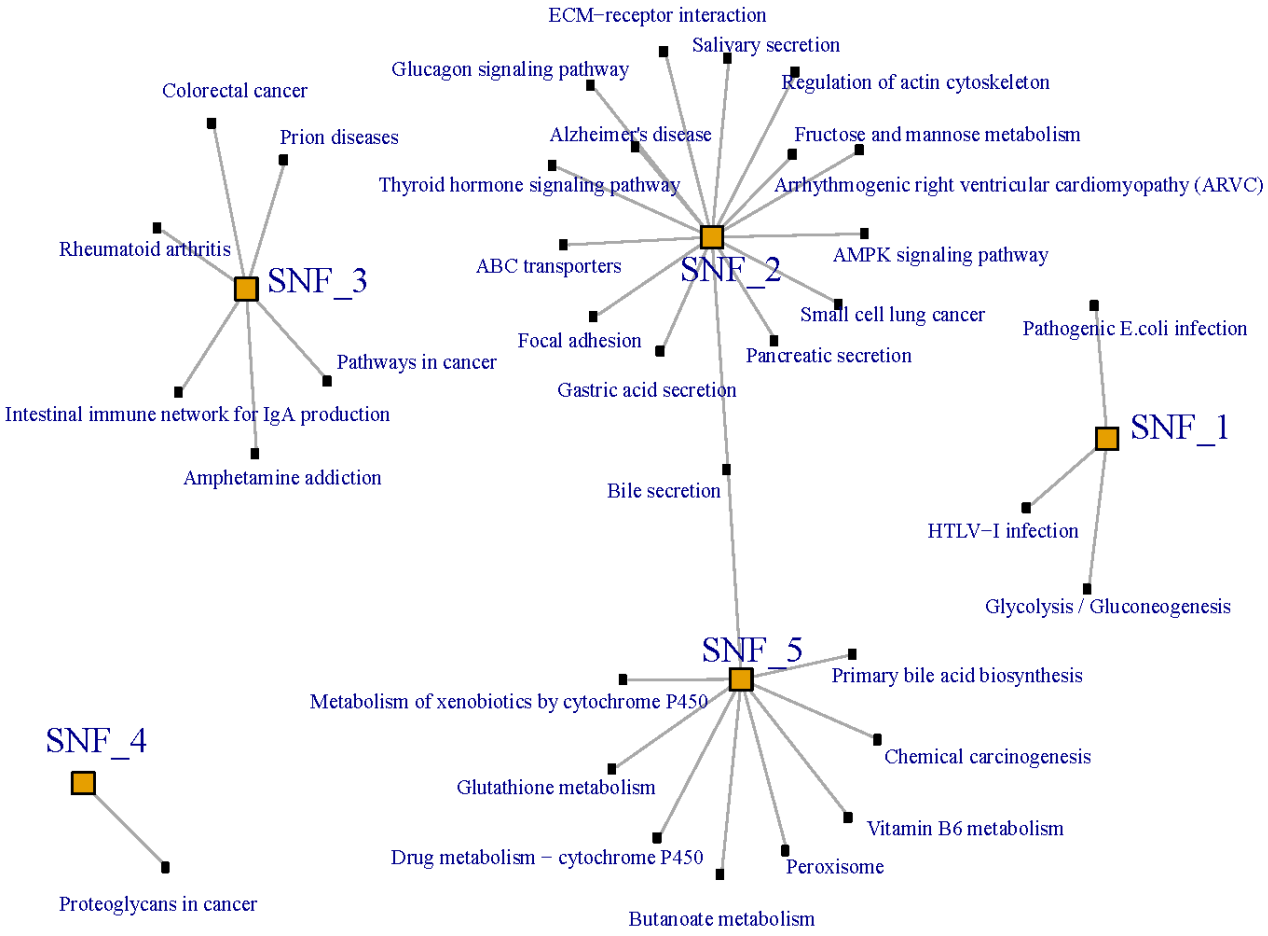

Figure 26: KEGG pathway enrichment in HCC subtypes with representative genes from gene expression

In order to further characterize the SNF subtypes, we used the class centroid approach described earlier in the predictive model training section, and selected the highly expressed genes for each subtype by centroids filtering. Furthermore, we conducted KEGG pathway enrichment analysis using package limma (Ritchie et al. 2015), and illustrate the significantly enriched pathways in each subtype (Figure 26). These subtypes have very different and (almost) disjoint active pathways, confirming that they are distinct subgroups at the molecular gene expression level. SNF_1 has active processes of cell cycle and DNA replication, including cell cycle, oocyte meiosis, p53 signaling pathway, DNA replication, homologous recombination and mismatch repair. SNF_2 has the most highly expressed genes and these genes are enriched in signaling pathways, such as hippo signaling pathway, cAMP signaling pathway, AMPK signaling pathway, hedgehog signaling pathway, PI3K-Akt signaling pathway and ECM-receptor interaction. SNF_3 also has active PI3K-Akt signaling pathway and ECM-receptor interaction. Additionally, SNF_3 has active metabolic processes, including cyanoamino acid metabolism, taurine and hypotaurine metabolism, and selenocompound metabolism. SNF_4 has active Phototransduction (RHO) and ABC transporters (ABCC10). As the genes in SNF_5 has lower expression compared to other classes, SNF_5 has deregulated metabolic processes, including metabolism of xenobiotics by cytochrome P450, butanoate metabolism, pyruvate metabolism, vitamin B6 metabolism, glutathione metabolism, primary bile acid biosynthesis

Chapter 5. Conclusions

5.1 Summary of results

In this thesis, we have performed a comprehensive study on the 13 putative driver genes in HCC using the TCGA tumor and tumor adjacent normal tissue exome-seq data. These putative driver genes are consistent with previous drive gene studies (M. Li et al. 2011; Guichard et al. 2012; J. Huang et al. 2012; Fujimoto et al. 2012; Cleary et al. 2013; Llovet et al. 2016). These putative driver genes cause their own gene expression changes, and are involved in important biological pathways by themselves. Moreover, we have discovered significant associations between clinical

characteristics and putative driver genes. In the survival association, these putative driver genes could effectively predict the risks on patients. Most importantly, we have presented that the putative driver genes have impacted about 44.5% gene expression, through building a multiple linear regression model between putative driver genes and gene expression. These influenced genes are involved in various pathways in proliferation, signaling transduction and metabolism. Similarly, we performed linear regression on miRNA expression, and we obtained 18 miRNAs whose expression is significantly impacted by the putative driver genes.

In the section of "Multi-omic data integration to stratify population in hepatocellular carcinoma", we used similarity network fusion approach to integrate multi-omics information and identified five subtypes from the molecular level. These subtypes have different combinations of driver gene mutations. And these subtypes are able to predict survival as being independent from other clinical histological features known to be related to survival. In the association with clinical characteristics, SNF_3 has the best survival, since they have similar profiles with normal samples. We further implemented the class centroids as features and built predictive models for new patients on each dataset. Although HCC is one of the most heterogeneous cancers, the predictive model on four datasets capture 57% to 87% consistency on individual omics level. This indicates the efficiency of noise-removal using the multi-omics approach.

5.2 Significance

In the analysis of Clinical and transcriptomics associations of putative driver mutations in hepatocellular carcinoma, we built survival model and linear regression model to elucidate the putative driver gene functions from the TCGA HCC cohort. We confirmed that TP53 and CTNNB1 are two most prevalent mutated genes, affecting 25-32% HCC patients in TCGA data. With these putative driver genes as features, we showed that they have significance on survival risks through Cox-PH modelling. We also discovered other genes ALB, BAP1 and APOB as being potential driver genes. About 45% genes' expression is significantly associated with the putative driver genes. Through pathway enrichment analysis, we observed that CTNNB1, BAP1, TP53 and RB1 are the key driver mutations with vast effects on HCC development.

In the molecular classification of HCC, we integrated four omic datasets and built sub-clusters in the consideration of complementary information from different levels of omic data. With the association of survival status, we obtained 5 optimal subtypes, which demonstrated significance on overall survival and relapse-free survival. These molecular subtypes could function as an independent feature to predict the survival in clinics. Furthermore, in the application of our molecular subtypes, we showed that single omics predictive model has 56-87% concordance with our fused omic data model. To our knowledge, this is the first time that we took advantage of comprehensive omics information to establish molecular classification and elucidate the related biological mechanism on HCC.

REFERENCES

Aguilar, Fernando, S. Perwez Hussain, and Peter Cerutti. 1993. "Aflatoxin B1 Induces the Transversion of G--> T in Codon 249 of the p53 Tumor Suppressor Gene in Human Hepatocytes." *Proceedings of the National Academy of Sciences* 90 (18). National Acad Sciences: 8586–90.

Ahn, Sung-Min, Se Jin Jang, Ju Hyun Shim, Deokhoon Kim, Seung-Mo Hong, Chang Ohk Sung, Daehyun Baek, et al. 2014. "Genomic Portrait of Resectable Hepatocellular Carcinomas: Implications of RB1 and FGF19 Aberrations for Patient Stratification." *Hepatology* 60 (6): 1972–82.

Ayub, Ambreen, Usman Ali Ashfaq, and Asma Haque. 2013. "HBV Induced HCC: Major Risk Factors from Genetic to Molecular Level." *BioMed Research International* 2013 (August): 810461.

Bartel, David P. 2004. "MicroRNAs: Genomics, Biogenesis, Mechanism, and Function." *Cell* 116 (2): 281–97.

Boyault, Sandrine, David S. Rickman, Aurélien de Reyniès, Charles Balabaud, Sandra Rebouissou, Emmanuelle Jeannot, Aurélie Hérault, et al. 2007. "Transcriptome Classification of HCC Is Related to Gene Alterations and to New Therapeutic Targets." *Hepatology* 45 (1): 42–52.

Bühler, Sandra, and Ralf Bartenschlager. 2012. "Promotion of Hepatocellular Carcinoma by Hepatitis C Virus." *Digestive Diseases* 30 (5): 445–52.

Calvisi, Diego F., Sara Ladu, Alexis Gorden, Miriam Farina, Ju-Seog Lee, Elizabeth A. Conner, Insa Schroeder, Valentina M. Factor, and Snorri S. Thorgeirsson. 2007. "Mechanistic and Prognostic Significance of Aberrant Methylation in the Molecular Pathogenesis of Human Hepatocellular Carcinoma." *The Journal of Clinical Investigation* 117 (9): 2713–22.

Cancer Genome Atlas Network. 2015. "Comprehensive Genomic Characterization of Head and Neck Squamous Cell Carcinomas." *Nature* 517 (7536): 576–82.

Chen, Mingquan, Jianbin Zhang, Ning Li, Zhiping Qian, Mengqi Zhu, Qian Li, Jianming Zheng, Xinyu Wang, and Guangfeng Shi. 2011. "Promoter Hypermethylation Mediated Downregulation of FBP1 in Human Hepatocellular Carcinoma and Colon Cancer." *PloS One* 6 (10): e25564.

Cleary, Sean P., William R. Jeck, Xiaobei Zhao, Kui Chen, Sara R. Selitsky, Gleb L. Savich, Ting-Xu Tan, et al. 2013. "Identification of Driver Genes in Hepatocellular Carcinoma by Exome Sequencing." *Hepatology* 58 (5): 1693–1702.

Cox, D. R. 1972. “Regression Models and Life-Tables.” *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 34 (2). [Royal Statistical Society, Wiley]: 187–220.
Dumitrescu, Ramona G. 2009. “Epigenetic Targets in Cancer Epidemiology.” *Methods in Molecular Biology* 471: 457–67.
El–Serag, Hashem B., and K. Lenhard Rudolph. 2007. “Hepatocellular Carcinoma: Epidemiology and Molecular Carcinogenesis.” *Gastroenterology* 132 (7): 2557–76.
Fan, H., L. Chen, F. Zhang, Y. Quan, X. Su, X. Qiu, Z. Zhao, et al. 2012. “MTSS1, a Novel Target of DNA Methyltransferase 3B, Functions as a Tumor Suppressor in Hepatocellular Carcinoma.” *Oncogene* 31 (18): 2298–2308.
Forbes, Simon A., David Beare, Prasad Gunasekaran, Kenric Leung, Nidhi Bindal, Harry Boutselakis, Minjie Ding, et al. 2015. “COSMIC: Exploring the World’s Knowledge of Somatic Mutations in Human Cancer.” *Nucleic Acids Research* 43 (Database issue): D805–11.
Friedler, Assaf, Brian S. DeDecker, Stefan M. V. Freund, Caroline Blair, Stefan Rüdiger, and Alan R. Fersht. 2004. “Structural Distortion of p53 by the Mutation R249S and Its Rescue by a Designed Peptide: Implications for ‘Mutant Conformation.’” *Journal of Molecular Biology* 336 (1): 187–96.
Fujimoto, Akihiro, Yasushi Totoki, Tetsuo Abe, Keith A. Boroevich, Fumie Hosoda, Ha Hai Nguyen, Masayuki Aoki, et al. 2012. “Whole-Genome Sequencing of Liver Cancers Identifies Etiological Influences on Mutation Patterns and Recurrent Mutations in Chromatin Regulators.” *Nature Genetics* 44 (7). nature.com: 760–64.
Gerstung, Moritz, Andrea Pellagatti, Luca Malcovati, Aristoteles Giagounidis, Matteo G. Della Porta, Martin Jädersten, Hamid Dolatshad, et al. 2015. “Combining Gene Mutation with Gene Expression Data Improves Outcome Prediction in Myelodysplastic Syndromes.” *Nature Communications* 6 (January): 5901.
Goossens, Nicolas, Xiaochen Sun, and Yujin Hoshida. 2015. “Molecular Classification of Hepatocellular Carcinoma: Potential Therapeutic Implications.” *Hepatic Oncology* 2 (4): 371–79.
Guichard, Cécile, Giuliana Amaddeo, Sandrine Imbeaud, Yannick Ladeiro, Laura Pelletier, Ichrafe Ben Maad, Julien Calderaro, et al. 2012. “Integrated Analysis of Somatic Mutations and Focal Copy-Number Changes Identifies Key Genes and Pathways in Hepatocellular Carcinoma.” *Nature Genetics* 44 (6). nature.com: 694–98.
Harrell, F. E., Jr, K. L. Lee, and D. B. Mark. 1996. “Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors.” *Statistics in Medicine* 15 (4): 361–87.
Hartwell, Hadley J., Keiko Y. Petrosky, James G. Fox, Nelson D. Horseman, and Arlin B. Rogers. 2014. “Prolactin Prevents Hepatocellular Carcinoma by Restricting Innate Immune Activation of c-Myc in Mice.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (31): 11455–60.
Hoshida, Yujin, Sebastian M. B. Nijman, Masahiro Kobayashi, Jennifer A. Chan, Jean-Philippe Brunet, Derek Y. Chiang, Augusto Villanueva, et al. 2009. “Integrative Transcriptome Analysis Reveals Common Molecular Subclasses of Human Hepatocellular Carcinoma.” *Cancer Research* 69 (18): 7385–92.
Huang, Jian, Qing Deng, Qun Wang, Kun-Yu Li, Ji-Hong Dai, Niu Li, Zhi-Dong Zhu, et al. 2012. “Exome Sequencing of Hepatitis B Virus-Associated Hepatocellular Carcinoma.” *Nature Genetics* 44 (10). nature.com: 1117–21.
Huang, Qichao, Biaoyang Lin, Hanqiang Liu, Xi Ma, Fan Mo, Wei Yu, Lisha Li, et al. 2011. “RNA-Seq Analyses Generate Comprehensive Transcriptomic Landscape and Reveal Complex Transcript Patterns in Hepatocellular Carcinoma.” *PloS One* 6 (10): e26168.
Huang, Sijia, Cameron Yee, Travers Ching, Herbert Yu, and Lana X. Garmire. 2014. “A Novel Model to Combine Clinical and Pathway-Based Transcriptomic Information for the Prognosis Prediction of

Breast Cancer." *PLoS Computational Biology* 10 (9): e1003851.
Hyun, Seogang, Jung Hyun Lee, Hua Jin, Jinwu Nam, Bumjin Namkoong, Gina Lee, Jongkyeong Chung, and V. Narry Kim. 2009. "Conserved MicroRNA miR-8/miR-200 and Its Target USH/FOG2 Control Growth by Regulating PI3K." *Cell* 139 (6): 1096–1108.
the International Cancer Genome Consortium Mutation Pathways and Consequences Subgroup of the Bioinformatics Analyses Working Group. 2013. "Computational Approaches to Identify Functional Genetic Variants in Cancer Genomes." *Nature Methods* 10 (8). Nature Publishing Group: 723–29.
Jiang, Runqiu, Lei Deng, Liang Zhao, Xiangcheng Li, Feng Zhang, Yongxiang Xia, Yun Gao, Xuehao Wang, and Beicheng Sun. 2011. "miR-22 Promotes HBV-Related Hepatocellular Carcinoma Development in Males." *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 17 (17): 5593–5603.
Ji, Junfang, Taro Yamashita, Anuradha Budhu, Marshonna Forgues, Hu-Liang Jia, Cuiling Li, Chuxia Deng, et al. 2009. "Identification of microRNA-181 by Genome-Wide Screening as a Critical Player in EpCAM–positive Hepatic Cancer Stem Cells." *Hepatology* 50 (2). Wiley Subscription Services, Inc., A Wiley Company: 472–80.
Joerger, Andreas C., Hwee Ching Ang, Dmitry B. Veprintsev, Caroline M. Blair, and Alan R. Fersht. 2005. "Structures of p53 Cancer Mutants and Mechanism of Rescue by Second-Site Suppressor Mutations." *The Journal of Biological Chemistry* 280 (16): 16030–37.
Kan, Zhengyan, Hancheng Zheng, Xiao Liu, Shuyu Li, Thomas D. Barber, Zhuolin Gong, Huan Gao, et al. 2013. "Whole-Genome Sequencing Identifies Recurrent Mutations in Hepatocellular Carcinoma." *Genome Research* 23 (9). genome.cshlp.org: 1422–33.
Laursen, Lucas. 2014. "A Preventable Cancer." *Nature* 516 (7529): S2–3.
Law, Charity W., Yunshun Chen, Wei Shi, and Gordon K. Smyth. 2014. "Voom: Precision Weights Unlock Linear Model Analysis Tools for RNA-Seq Read Counts." *Genome Biology* 15 (2): R29.
Lawrence, Michael S., Petar Stojanov, Paz Polak, Gregory V. Kryukov, Kristian Cibulskis, Andrey Sivachenko, Scott L. Carter, et al. 2013. "Mutational Heterogeneity in Cancer and the Search for New Cancer-Associated Genes." *Nature* 499 (7457). nature.com: 214–18.
Li, Meng, Hong Zhao, Xiaosong Zhang, Laura D. Wood, Robert A. Anders, Michael A. Choti, Timothy M. Pawlik, et al. 2011. "Inactivating Mutations of the Chromatin Remodeling Gene ARID2 in Hepatocellular Carcinoma." *Nature Genetics* 43 (9). nature.com: 828–29.
Li, Yvonne Y., Glenn J. Hanna, Alvaro C. Laga, Robert I. Haddad, Jochen H. Lorch, and Peter S. Hammerman. 2015. "Genomic Analysis of Metastatic Cutaneous Squamous Cell Carcinoma." *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 21 (6): 1447–56.
Llovet, Josep M., Jessica Zucman-Rossi, Eli Pikarsky, Bruno Sangro, Myron Schwartz, Morris Sherman, and Gregory Gores. 2016. "Hepatocellular Carcinoma." *Nature Reviews. Disease Primers* 2 (April): 16018.
Maass, Thorsten, Ioannis Sfakianakis, Frank Staib, Markus Krupp, Peter R. Galle, and Andreas Teufel. 2010. "Microarray-Based Gene Expression Analysis of Hepatocellular Carcinoma." *Current Genomics* 11 (4): 261–68.
Meng, Fanyin, Roger Henson, Hania Wehbe-Janek, Kalpana Ghoshal, Samson T. Jacob, and Tushar Patel. 2007. "MicroRNA-21 Regulates Expression of the PTEN Tumor Suppressor Gene in Human Hepatocellular Cancer." *Gastroenterology* 133 (2): 647–58.
Morishita, Asahiro, and Tsutomu Masaki. 2015. "miRNA in Hepatocellular Carcinoma." *Hepatology Research: The Official Journal of the Japan Society of Hepatology* 45 (2): 128–41.
Rebouissou, Sandra, Andrea Franconi, Julien Calderaro, Eric Letouzé, Sandrine Imbeaud, Camilla Pilati, Jean-Charles Nault, et al. 2016. "Genotype-Phenotype Correlation of CTNNB1 Mutations Reveals Different ß-Catenin Activity Associated with Liver Tumor Progression." *Hepatology* , May.

doi:10.1002/hep.28638.
Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. “Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies.” *Nucleic Acids Research* 43 (7): e47.
Schlaeger, Christof, Thomas Longerich, Claudia Schiller, Peter Bewerunge, Arianeb Mehrabi, Grischa Toedt, Jörg Kleeff, et al. 2008. “Etiology-Dependent Molecular Mechanisms in Human Hepatocarcinogenesis.” *Hepatology* 47 (2): 511–20.
Schröder, Markus S., Aedín C. Culhane, John Quackenbush, and Benjamin Haibe-Kains. 2011. “Survcomp: An R/Bioconductor Package for Performance Assessment and Comparison of Survival Models.” *Bioinformatics* 27 (22): 3206–8.
Schulze, Kornelius, Sandrine Imbeaud, Eric Letouzé, Ludmil B. Alexandrov, Julien Calderaro, Sandra Rebouissou, Gabrielle Couchy, et al. 2015. “Exome Sequencing of Hepatocellular Carcinomas Identifies New Mutational Signatures and Potential Therapeutic Targets.” *Nature Genetics* 47 (5): 505–11.
Seton-Rogers, Sarah. 2014. “Hepatocellular Carcinoma: Gender Differences.” *Nature Reviews. Cancer* 14 (9): 578.
Shibata, Tatsuhiro, and Hiroyuki Aburatani. 2014. “Exploration of Liver Cancer Genomes.” *Nature Reviews. Gastroenterology & Hepatology* 11 (6): 340–49.
Siegel, Rebecca L., Kimberly D. Miller, and Ahmedin Jemal. 2015. “Cancer Statistics, 2015.” *CA: A Cancer Journal for Clinicians* 65 (1). Wiley Online Library: 5–29.
Song, Min-Ae, Maarit Tiirikainen, Sandi Kwee, Gordon Okimoto, Herbert Yu, and Linda L. Wong. 2013. “Elucidating the Landscape of Aberrant DNA Methylation in Hepatocellular Carcinoma.” *PloS One* 8 (2): e55761.
Testino, Gianni, Silvia Leone, and Paolo Borro. 2014. “Alcohol and Hepatocellular Carcinoma: A Review and a Point of View.” *World Journal of Gastroenterology: WJG* 20 (43): 15943–54.
Therneau, Terry M., and Patricia M. Grambsch. 2000. *Modeling Survival Data: Extending the Cox Model:* Statistics for Biology and Health . Springer New York.
Tibshirani, Robert, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. 2002. “Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression.” *Proceedings of the National Academy of Sciences of the United States of America* 99 (10): 6567–72.
Tomimaru, Yoshito, Hidetoshi Eguchi, Hiroaki Nagano, Hiroshi Wada, Shogo Kobayashi, Shigeru Marubashi, Masahiro Tanemura, et al. 2012. “Circulating microRNA-21 as a Novel Biomarker for Hepatocellular Carcinoma.” *Journal of Hepatology* 56 (1). Elsevier: 167–75.
Torre, Lindsey A., Freddie Bray, Rebecca L. Siegel, Jacques Ferlay, Joannie Lortet-Tieulent, and Ahmedin Jemal. 2015. “Global Cancer Statistics, 2012.” *CA: A Cancer Journal for Clinicians* 65 (2): 87–108.
Totoki, Yasushi, Kenji Tatsuno, Kyle R. Covington, Hiroki Ueda, Chad J. Creighton, Mamoru Kato, Shingo Tsuji, et al. 2014. “Trans-Ancestry Mutational Landscape of Hepatocellular Carcinoma Genomes.” *Nature Genetics* 46 (12): 1267–73.
Totoki, Yasushi, Kenji Tatsuno, Shogo Yamamoto, Yasuhito Arai, Fumie Hosoda, Shumpei Ishikawa, Shuichi Tsutsumi, et al. 2011. “High-Resolution Characterization of a Hepatocellular Carcinoma Genome.” *Nature Genetics* 43 (5). nature.com: 464–69.
Tsai, Chia-Chu, Kai-Wen Huang, Hsiao-Fen Chen, Bo-Wen Zhan, Yen-Han Lai, Fa-Han Lee, Chung-Yei Lin, et al. 2006. “Gene Expression Analysis of Human Hepatocellular Carcinoma by Using Full-Length cDNA Library.” *Journal of Biomedical Science* 13 (2). Springer Netherlands: 241–49.
Vandin, Fabio, Eli Upfal, and Benjamin J. Raphael. 2012. “De Novo Discovery of Mutated Driver Pathways in Cancer.” *Genome Research* 22 (2): 375–85.
Vogelstein, Bert, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz Jr, and

Kenneth W. Kinzler. 2013. “Cancer Genome Landscapes.” *Science* 339 (6127). sciencemag.org: 1546–58.
Wang, Bo, Aziz M. Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. 2014. “Similarity Network Fusion for Aggregating Data Types on a Genomic Scale.” *Nature Methods* 11 (3): 333–37.
Wang, Kai, Ho Yeong Lim, Stephanie Shi, Jeeyun Lee, Shibing Deng, Tao Xie, Zhou Zhu, et al. 2013. “Genomic Landscape of Copy Number Aberrations Enables the Identification of Oncogenic Drivers in Hepatocellular Carcinoma.” *Hepatology* 58 (2): 706–17.
Woo, Hyun Goo, Eun Sung Park, Ju-Seog Lee, Yun-Han Lee, Tsuyoshi Ishikawa, Yoon Jun Kim, and Snorri S. Thorgeirsson. 2009. “Identification of Potential Driver Genes in Human Liver Carcinoma by Genomewide Screening.” *Cancer Research* 69 (9): 4059–66.
Wurmbach, Elisa, Ying-Bei Chen, Greg Khitrov, Weijia Zhang, Sasan Roayaie, Myron Schwartz, Isabel Fiel, et al. 2007. “Genome-Wide Molecular Profiles of HCV-Induced Dysplasia and Hepatocellular Carcinoma.” *Hepatology* 45 (4): 938–47.
Xiang, Qian, Xianhua Dai, Yangyang Deng, Caisheng He, Jiang Wang, Jihua Feng, and Zhiming Dai. 2008. “Missing Value Imputation for Microarray Gene Expression Data Using Histone Acetylation Information.” *BMC Bioinformatics* 9 (May): 252.
Xie, Boya, Qin Ding, Hongjin Han, and Di Wu. 2013. “miRCancer: A microRNA-Cancer Association Database Constructed by Text Mining on Literature.” *Bioinformatics* 29 (5): 638–44.
Yan, Shi-Yan, Mei-Mei Chen, Guang-Ming Li, Yu-Qin Wang, and Jian-Gao Fan. 2015. “MiR-32 Induces Cell Proliferation, Migration, and Invasion in Hepatocellular Carcinoma by Targeting PTEN.” *Tumour Biology: The Journal of the International Society for Oncodevelopmental Biology and Medicine* 36 (6): 4747–55.
You, Xiaona, Fabao Liu, Tao Zhang, Yinghui Li, Lihong Ye, and Xiaodong Zhang. 2013. “Hepatitis B Virus X Protein Upregulates Oncogene Rab18 to Result in the Dysregulation of Lipogenesis and Proliferation of Hepatoma Cells.” *Carcinogenesis* 34 (7): 1644–52.
Yu, Guangchuang, Li-Gen Wang, Yanyan Han, and Qing-Yu He. 2012. “clusterProfiler: An R Package for Comparing Biological Themes among Gene Clusters.” *Omics: A Journal of Integrative Biology* 16 (5): 284–87.
Zhang, Xiaoying, Hiu Ming Li, Zhiyan Liu, Gengyin Zhou, Qinghui Zhang, Tingguo Zhang, Jianping Zhang, and Cuijuan Zhang. 2013. “Loss of Heterozygosity and Methylation of Multiple Tumor Suppressor Genes on Chromosome 3 in Hepatocellular Carcinoma.” *Journal of Gastroenterology* 48 (1): 132–43.
Zhao, Zhujiang, Qingxiang Wu, Jian Cheng, Xuemei Qiu, Jianqiong Zhang, and Hong Fan. 2010. “Depletion of DNMT3A Suppressed Cell Proliferation and Restored PTEN in Hepatocellular Carcinoma Cell.” *Journal of Biomedicine & Biotechnology* 2010 (May): 737535.
Zhu, Yitan, Peng Qiu, and Yuan Ji. 2014. “TCGA-Assembler: Open-Source Software for Retrieving and Processing TCGA Data.” *Nature Methods* 11 (6): 599–600.
Zucman-Rossi, Jessica, Augusto Villanueva, Jean-Charles Nault, and Josep M. Llovet. 2015. “Genetic Landscape and Biomarkers of Hepatocellular Carcinoma.” *Gastroenterology* 149 (5): 1226–39.e4.