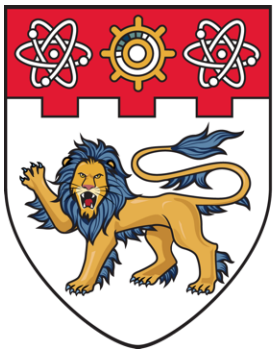


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

An evolutionary and genomic approach to understand tumor evolution in hepatocellular carcinoma

Arife Neslihan Kaya

SCHOOL OF BIOLOGICAL SCIENCES

2021

An evolutionary and genomic approach to understand tumor evolution in hepatocellular carcinoma

Arife Neslihan Kaya

SCHOOL OF BIOLOGICAL SCIENCES

A thesis submitted to the Nanyang Technological University in partial fulfilment of the requirement for the degree of Doctor of Philosophy

2021

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research done by me except where otherwise stated in this thesis. The thesis work has not been submitted for a degree or professional qualification to any other university or institution. I declare that this thesis is written by myself and is free of plagiarism and of sufficient grammatical clarity to be examined. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

12/01/2021

. . . . . .

Date

. . . . . .

Arife Neslihan Kaya

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it of sufficient grammatical clarity to be examined. To the best of my knowledge, the thesis is free of plagiarism and the research and writing are those of the candidate's except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

13/01/2021

. . . . . .
Date

. . . . . .
Assoc Prof Mu Yuguang

Authorship Attribution Statement

This thesis **does not** contain any materials from papers published in peer-reviewed journals or from papers accepted at conferences in which I am listed as an author.

12/01/2021

. . . . . .

Date

. . . . . .

Arife Neslihan Kaya

Acknowledgements

First of all, I would like to thank my supervisors Assoc. Prof. Mu Yuguang and Dr. Torsten Wuestefeld for their supervision. I appreciate their contributions during the course of my Ph.D. project as well as for their continuous support during the uncertain times of my Ph.D. including a co-supervisor change. I also would like to express my endless gratitude to Dr. Zhai Weiwei, for being such a great supervisor/mentor before and after leaving Singapore in the second year of my PhD. His supports and close guidance for my projects helped me gain required resistance and skills to finish my Ph.D. I would also like to acknowledge my thesis advisory committee member Prof. Zbynek Bozdech from the Nanyang Technological University and Prof. Pierce Chow, the director of the PLANET project, for providing us clinical samples and sharing his clinical vision with our team.

I would like to thank my colleagues in the Genome Institute of Singapore, in particular my team friends who could manage to stay as a group after our PI leaves. Special thanks to Jia Qi Lim, our former teammate, for conducting the wet-lab experiments required to generate needed data in the part of my thesis, also for being a close friend to me for four years and teaching me about the Singaporean culture as well as a little bit Chinese. I also would also like to thank Dr. Jianbin Chen for keeping our team together and for his stimulating discussions.

In addition, I would like to thank my father Cihat, my mother Hatice as well as my siblings Nedim and Damla for supporting me and believing in me. Their

efforts to keep me happy from thousands of miles away helped me stay positive at hard times. I would like to thank my special friend, a gift I have gained from my Ph.D. journey, Anil Kurkcu for his love and endless support during this challenging period. Last, but not least, I would like to thank my close friends Ezgi Burgazdere and Ceylan Aydin for their kind supports.

Dedication

This thesis is dedicated to the memory of Sait Birol Kurkcu (1961-2019).

Contents

List of Figures

List of Tables

Abbreviations

Table 1: Table of abbreviations

Abbreviation	Definition
AFP	Alpha-fetoprotein
AMP	Amplification
BAM	Binary alignment map
CCF	Cancer cell fraction
CGC	Cancer gene census
Chr	Chromosome
CIN	Copy number instability
CN	Copy number
CNV	Copy number variation
COSMIC	Catalogue Of Somatic Mutations In Cancer
DEL	Deletion
FDR	False discovery rate
GD	Genome doubling
GDC	Genomic data commons
GII	Genomic instability index
GISTIC	Genomic Identification of Significant Targets in Cancer
GSVA	Gene set variation analysis
HBV	Hepatitis B virus
HCC	Hepatocellular carcinoma
HCV	Hepatitis C virus

Abbreviation	Definition
ICGC	International Cancer Genome Consortium
ICI	Immune checkpoint inhibitor
ITH	Intratumor heterogeneity
KEGG	Kyoto Encyclopedia of Genes and Genomes
LRT	Likelihood ratio test
MATH	Mutant allele heterogeneity
MCMC	Markov chain Monte Carlo
MDSC	Myeloid derived suppressor cell
MSI	Microsatellite instability
MVI	Microvascular invasion
NMF	Non-negative matrix factorization
OG	Oncogene
OR	Odds ratio
PC	Principle component
PLANET	Precision medicine in Liver cancer across an Asia-pacific NETwork
pLM	Percentage of late mutations
RIKEN	Institute of Physical and Chemical Research
SBS	Single base substitution
SCNA	Somatic copy number alteration
SG	Signature group
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variation
TAM	Tumor associated macrophage

Abbreviation	Definition
TCGA	The Cancer Genome Atlas
TIDE	Tumor Immune Dysfunction and Exclusion
TIL	Tumor infiltrating lymphocyte
TMB	Tumor mutation burden
TSG	Tumor Suppressor Gene
TUSON	Tumor Suppressor and Oncogene
UTR	Untranslated region
VAF	Variant allele frequency
WES	Whole exome sequencing
WGS	Whole genome sequencing
WT	Wild type

Abstract

Hepatocellular carcinoma (HCC) is one of the deadliest cancer types with diverse etiological factors across the world and very limited treatment options. Although large-scale genomic and transcriptomic studies have been conducted in different cohorts, an integrative analysis of HCC genomes and the ethnic comparison across ethnic backgrounds is lacking. In the second chapter, we integrated 1349 HCC genomes from five Asian and/or European cohorts and identified a large number of novel drivers (n=29) (e.g. *FRG1*). Many of these novel drivers are infrequent tumor suppressor genes (TSGs) and tend to enrich in several important pathways including the chromatin remodeling pathway. Novel drivers including *PBRM1* and *KMT2D* often occur in later stages of tumorigenesis indicating their potential roles in driving tumor progression.

In the third chapter, in order to understand ethnic differences in HCC, we conducted a systematic comparison across the Asian and European HCCs using the data from the The Cancer Genome Atlas (TCGA) database. We found higher genomic instability in Asians with a series of molecular events ranging from driver genes to immune profiles segregating distinctively between the two ethnic backgrounds. Most strikingly, multiple Asian enriched genomic alterations in particular chromosome 16 deletion, lead to a clinically aggressive RNA subgroup unique to Asians. By integrating information across multiple layers, we found that integrative survival models predict patient prognosis much better in Asians, suggesting a better chance to conduct a precision medicine program in Asia. Taken together, we performed a comprehensive integrative

analysis of HCC genomes and uncovered remarkable ethnic differences between Asians and Europeans.

In the fourth chapter of the thesis, leveraging the multi-region data of an Asian cohort, PLANET, variable level of intra-tumor heterogeneity (ITH) was found in both genomes and transcriptomes of HCC. Strikingly, we found multiple RNA subtypes within a single patient in the PLANET cohort which might be one reason for the poor treatment response in HCC. Integrative survival analysis highlighted the important roles of DNA and RNA ITH in predicting progression-free patient survival. Taken together, we have drawn a comprehensive landscape of ITH across a prospective cohort for HCC.

1 Chapter 1: Introduction

1.1 Introduction

Cancer is a genetic disease caused by uncontrollable proliferation of malignant cells (Hanahan & Weinberg, 2000, 2011). Starting from the initial observation on chromosomal abnormalities in cancer cells more than a century ago, understanding genetic changes leading to cancer has been a major endeavor for the field of genetics (Boveri, 2008; Manchester, 1995). With the rise of genomic technologies, a large number of cancer genomes have been extensively characterized (Deng et al., 2017; Ding et al., 2018; Goodwin et al., 2016; Loveday et al., 2020; Mäkinen et al., 2016; Wang et al., 2020).

Despite many years of effort, there are still many gaps in cancer genomics. First, when somatic mutations are identified in a cancer genome, it includes mutations accumulated during early development as well as carcinogenesis. Identifying driver genomic alterations and distinguishing them from insignificant ones (so called passenger mutations) is a challenge in this area (Chin et al., 2011; Hofree et al., 2016; Stratton, 2011). Secondly, even though hallmarks of cancer have been described for many years, due to the pleiotropic effect of genes, understanding the functional consequence of all driver genes in a coherent and functional perspective is still difficult (Bien & Peters, 2019; Wu et al., 2018). Finally, genetic heterogeneity in tumors is one of the major barriers to comprehensive characterization of the disease and is a big roadblock to effective cancer therapies (Bozic & Wu, 2020; Jamal-Hanjani et al., 2015; McGranahan et al., 2015). Tackling these challenges in coming genomic studies are likely

to pave important basis for deeper understanding of cancer and develop more effective therapeutics against this deadly disease.

1.1.1 Hepatocellular carcinoma

Hepatocellular carcinoma (HCC) is the major subtype of liver cancer. It is the fifth most common cancer type overall and the second cause of cancer deaths worldwide (Bray et al., 2018). The risk factors for HCC include Hepatitis B & C virus infection, aflatoxin exposure and alcohol intake, as well as metabolic factors such as obesity and diabetes (Forner et al., 2018). Although several studies have been conducted to characterize the genomics of hepatocellular carcinoma, these studies were limited to individual cohorts with single ethnicity and etiological type (Ahn et al., 2014; Fujimoto et al., 2016; Guichard et al., 2012; Huang et al., 2012). Despite many years of efforts, systemic therapy for HCC is still very limited and Sorafenib, the only first-line targeted drug approved by FDA is beneficial only for a small proportion of HCC patients. Moreover, sorafenib prolongs overall survival by only 3 months (Llovet et al., 2008). Based on these facts, HCC remains as one of the deadliest cancer types, especially for Asia.

1.1.2 Rapid progress in cancer genomics

Since the rise of second-generation sequencing, sequencing technologies have allowed researchers to sequence whole human genomes, exomes (protein coding regions of the genome) and custom targeted regions effectively. These

technologies have rapidly transformed the speed and cost of interrogating the genetic information across many diseases, including cancer (Metzker, 2010).

Since then, cancer genomes have been extensively studied throughout individual groups as well as big international projects. For example, International Cancer Genome Consortium (ICGC) (Junjun Zhang et al., 2011) and The Cancer Genome Atlas (TCGA) (https://www.cancer.gov/tcga) have carried out genome sequencing of a large number of tumor types. As of today, genomic data for around 20,000 cases from 33 primary sites in TCGA and 22,330 cases from 22 primary cancer sites in ICGC are publicly available to researchers. Thus, a lot of cancer genomes have been characterized for HCC across the world.

1.1.3 Large scale studies of liver cancer genomes and known HCC drivers

Given the diverse etiologies of HCC, several countries and research groups have initiated their own efforts characterizing HCC genomes including Totoki et al. (2014) (Totoki et al., 2014) (Japanese cohorts, n=503), Ahn et al. (2014, Korean cohort, n=231) (Ahn et al., 2014), Schulze et al. (2015) (Schulze et al., 2015) (French cohort, n=243) and by The Cancer Genome Atlas Research Network (2017) (TCGA) (Ally et al., 2017) (US cohort, n=199). These large-scale studies have unveiled several driver genes and pathways associated with HCC.

One of the most frequently altered driver gene for HCC is *CTNNB1* which encodes for the protein β-catenin (Ahn et al., 2014; Ally et al., 2017; Totoki et

al., 2014). *TP53* is the other most frequently altered gene in HCC that codes for p53 tumor suppressor (C. Chen & Wang, 2015; Totoki et al., 2014). Several other driver genes including *ALB*, *APOB* and *AXIN1* were also reported by several large cohort studies. For example, Totoki et al. (2014) (Totoki et al., 2014) reported 30 drivers and a study by TCGA reported 26 driver genes (Ally et al., 2017).

In addition to understanding of drivers at the level of individual genes, genetic changes can also be categorized into pathways. For example, Wnt/β-catenin and p53 pathways have been reported to be the most frequently altered pathways in HCC (Cleary et al., 2013; Jhunjhunwala et al., 2014; Kan et al., 2013). *CTNNB1* in the canonical Wnt/β-catenin pathway has been reported to have activating mutations while the other driver *AXIN1* is acting as a suppressor by taking part in β-catenin destruction, had recurrent loss-of-function mutations in the same pathway (C. Chen & Wang, 2015). It has also been shown in several studies that recurrent mutations in chromatin remodeling pathway genes are frequently observed in HCC genomes (Guichard et al., 2012; Totoki et al., 2011). Both *ARID2* and *ARID1A* are subunits of the SWI/SNF (SWItch/Sucrose Non-Fermentable) complex (Wilson & Roberts, 2011). Genes from other signaling pathways such as *NFE2L2* and *KEAP1* in an oxidative stress pathway (Cleary et al., 2013), *CDKN2A* in a cell cycle pathway were also reported in several HCC studies (Guichard et al., 2012; Schulze et al., 2015). Thus, a large number of known drivers and pathways were found to be mutated in HCC genomes.

1.1.4 A short review of HCC molecular subtypes

High throughput molecular subtype characterization of HCC starts with profiling the transcriptome of 91 HCC patients and results in two distinct molecular subtypes (A and B, Figure 1.1) that are strongly associated with overall survival of patients (Lee et al., 2004). This study includes normal samples and gradual decrease in genes related to normal hepatocyte function from normal tissue to the poor survival subtype A is observed. Also, cell cycle pathway was found to be overexpressed in the poor survival subtype A (Figure 1.1). In a subsequent study by Lee et al. (2006), co-clustering of 61 Asian patients with fetal cells from the rat using the orthologous genes revealed that around 20% of human samples clustered together with fetal cells, indicating a potential evidence for emergence of HCC from progenitor cells (Figure 1.1).

Boyault et al. (2007) identified six (G1-G6) molecular subtypes using microarray data. This study is one of the first studies to show genetic associations of transcriptomic subtypes. Among these six subtypes, G1/G2/G3 showed activation of cell cycle pathways indicating a proliferating phenotype similar to poor survival subtype in Lee et al. (2004) (Figure 1.1). Importantly, these proliferative subtypes have a more unstable genome compared to the other subtypes (G4/G5/G6). They also showed that, G5 and G6 have overexpression of Wnt pathway and mutations in *CTNNB1* gene which is one of the most frequent HCC drivers (~30%) and a key player of Wnt pathway (Shibata & Aburatani, 2014). Interestingly, one of the proliferative subtypes (G1) showed quite unique expression signature with overexpression of genes

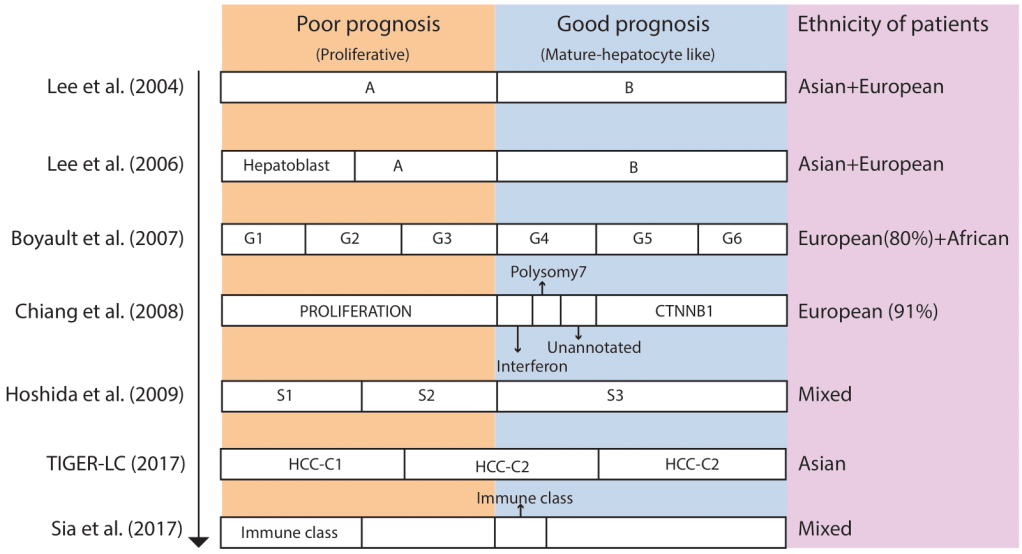


Figure 1.1: Chronological summary of HCC subtypes and ethnicities of patients. Major studies on HCC molecular subtypes were chronologically summarized. Subtypes were aligned according to their correspondence status based on comparisons in individual studies or obvious feature similarities. Ethicities of patients in each analysis cohort were also listed

expressed by fetal liver such as AFP, *PEG3* and *PEG10*. They also noted a link between G1 subtype and *AXIN1* mutations, another frequent HCC driver gene.

In 2008, Chiang et al. (2008) conducted an integrative analysis of 103 HCV related HCC patients. They identified five distinct transcriptomic clusters and tried to link each of them to genomics features of tumors using copy number data. Their subtpes named PROLIFERATION, matched to poor prognosis subtype identified by Lee et al. (2004) and G1/G2/G3 from Boyault et al. (2007). The rest of the subtypes were (INTERFERON, POLYSOMY7, CTNNB1, UNANNOTATED) similar to good prognosis subtypes (Figure 1.1). Notably, CTNNB1 subtype was similar to G5-G6 from Boyault et al. (2007).

In 2009, Hoshida et al. (2009) identified three HCC subtypes where two of them were poor prognosis (S1, S2) and the other one showed features of well differentiated hepatocyte features similar to good prognosis subtype identified by many previous studies (Figure 1.1). Chaisaingmongkol et al. (2017) identified three subtypes in a Thai cohort (C1-C3). C2 and C3 subtypes have better overall survival compared to C1. However, their comparison with Hoshida's subtypes shows that one of the good prognosis subtypes (C2) showed features of both good and bad prognosis subtypes (Figure 1.1). More recently, an immune subclass of HCC was identified by deconvoluting RNAseq data which mostly overlaps with the Hoshida's S1 subtype and a part of S3 as well as Chiang's INTERFERON subtype (Sia et al., 2017). In summary, while HCC subtypes reported in the literature often have two main subtypes as proliferative and differentiated hepatocyte-like, the classification if often very

different, possibly due to different etiologies and ethnicities across the cohorts.

1.1.5 Intra-Tumor heterogeneity

When mutations accumulate over the course of tumor initiation and development, distinct subpopulations of cells can co-exist within a single tumor and this phenomenon is often referred to as intra-tumor heterogeneity (ITH) (Marusyk et al., 2012) (Figure 1.2). At any time in tumorigenesis, a given tumor often comprises of a mixture of various cell populations and distinct subclones will tend to have different phenotypic properties possibly driven by subclonal mutations. While distinct subpopulations compete or collaborate to sustain their growth advantage, tumor heterogeneity is one of the biggest challenges hampering effective therapies for cancer. Developing methods to characterize ITH, understanding its origin and functional consequence are very important for the field (Greaves, 2015; McGranahan & Swanton, 2017).

The survey of tumor heterogeneity can be broadly classified into two subgroups. In the first approach, a single biopsy of the tumor is used and the clonal composition of the single biopsies are deconvoluted into subclones (Maley et al., 2017). Using this approach, tumor heterogeneity is found to be strongly linked to patient survival and stratification in several cancer types (Andor et al., 2016)

Even though a single biopsy can be deeply sequenced to detect rare subclones and can potentially capture genomic properties of the local population, biopsy itself might not represent the entire tumor as subclones can be localized to different locations of a tumor (Figure 1.2). To mitigate this limitation, the

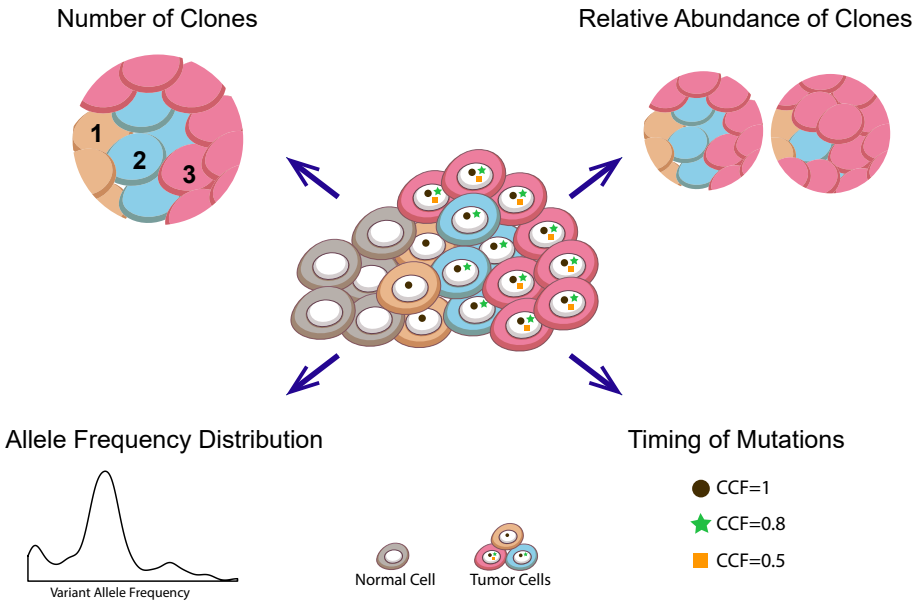


Figure 1.2: Metrics of intratumor heterogeneity. Different methods to measure intratumor heterogeneity (ITH) are illustrated. ITH can be measured by simply counting the number of distinct clones with distinct set of mutations (top left), relative abundance of clones (top right), with a dispersion metric derived from the variant allele frequency (VAF) distribution (bottom left) or by timing individual mutations based on their cancer cell fractions (CCF) which represents the fraction of cancer cells with the mutation of interest (bottom right). If the CCF equals 1, the mutation is present across all tumor cells and this mutation is likely to be an early mutation through the history of tumorigenesis.

other category of approaches surveys tumor heterogeneity using multi-sectoring approach by taking multiple sectors of the tumor (Gerlinger et al., 2012, 2014; Harbst et al., 2016; Jianjun Zhang et al., 2014). By sequencing multiple sectors instead of a single biopsy from a tumor, multi-region sequencing has discovered that spatial heterogeneity exists in several tumor types including lung adenocarcinoma (Jianjun Zhang et al., 2014), esophageal squamous cell carcinoma (Hao et al., 2016), clear cell renal carcinoma (Gerlinger et al., 2014), primary breast cancer (Yates et al., 2017) and melanoma (Harbst et al., 2016). Deconvolution of tumor heterogeneity and linking tumor heterogeneity to patient clinical trajectories and treatment response is an important topic in cancer genomics.

1.2 Thesis organization

My thesis consists of four chapters. Chapter 1 introduces important background information for the thesis including genomic drivers of HCC, a historical review of HCC molecular subtypes as well as the intratumor heterogeneity (ITH). The work presented in chapter 2 started in 2017 when many studies of HCC genomes have been conducted with medium size cohorts from different ethnic backgrounds, but integrative analysis across cohorts haven't been explored. In chapter 2, we conducted an integrative analysis of 1349 HCC genomes and depicted a comprehensive landscape of HCC genomes ranging from novel drivers and their functional pathways as well as the evolutionary timing of all the genomic changes. With increased statistical power, we were able to identify a larger number of novel drivers for HCC and characterized the evolutionary

history of all the genomic changes, providing a comprehensive overview of HCC genomes.

As HCC shows diverse geographical distribution and different disease etiologies, in Chapter 3, a systematic comparison of Asian and European HCC genomes is carried out leveraging the uniform platform provided by the TCGA cohort. We systematically compared the two cohorts across the clinical, genomic and transcriptomic levels. Most strikingly, an Asian specific aggressive RNA subtype and its potential genomic origin were explored and presented. Moreover, we found that integrative survival models predict patient survival much better in Asians than Europeans possibly driven by many ethnic specific events. For the first time, we characterized an unprecedented amount of ethnic differences across two ethnic backgrounds.

In Chapter 4, I presented a related work where I contributed to the analysis of a prospective cohort in HCC. The Precision medicine in Liver cancer across an Asia-pacific NETwork (PLANET) funded by Singapore National Medical Research Council is a prospective study that samples resected HCC from multi-ethnic sites within the established Asia-Pacific Hepatocellular Carcinoma (AHCC) Trials Group. PLANET aims to capture the natural history of all the patients from time of surgery to recurrence to understand how intra-tumor heterogeneity (ITH) might affect the clinical trajectories of HCC patients. In this Chapter, I reported on the initial findings from whole-genome sequencing (WGS) as well as RNA sequencing of the first 67 patients.

Even though, the three major chapters touched on slightly different topics in HCC, the coherent theme relating them is the genomic and evolutionary

landscape of hepatocellular carcinoma.

2 Chapter 2: An integrative approach to identify drivers for hepatocellular carcinoma

2.1 Introduction

The common approach to decode the cancer genome of a patient is to obtain samples from the tumor as well as the normal tissue (e.g., blood or the adjacent normal tissues) of the individual and sequencing them. One of the biggest challenges is understanding the functional consequence of all the somatic changes in the cancer genome. Somatic mutations accumulate in the genome due to either extrinsic factors (e.g. smoking or ultraviolet light exposure) or intrinsic factors (e.g. mutations accumulated during cell divisions in morphogenesis and tissue renewal). Only a few of these mutations give cells growth advantage over their neighboring cells and the ability to evade control mechanisms of the cell (Hanahan & Weinberg, 2000, 2011; Maley et al., 2017). Thus, cancer cells carry a large number of mutations and the minority of these mutations confer a fitness advantage for the cancer cells. Genes that can drive tumor initiation or progression are called drivers. The rest can be referred to as passengers. Distinguishing driver genes from passengers is a major problem in cancer genomics.

The identification of driver mutations requires a rigorous statistical analysis of mutations in a considerable size of cancer patients. In the past decade, several computational tools were developed for the identification of driver genes, each considering a distinct aspect of driver genes. Approaches taken so far

include such as 1) significantly higher frequency of mutations compared to the background mutation rate, 2) functional impact of mutations in a given gene as well as 3) the physical location and clustering of somatic mutations within a gene (e.g. oncogenes tend to have repeated changes in the activating sites).

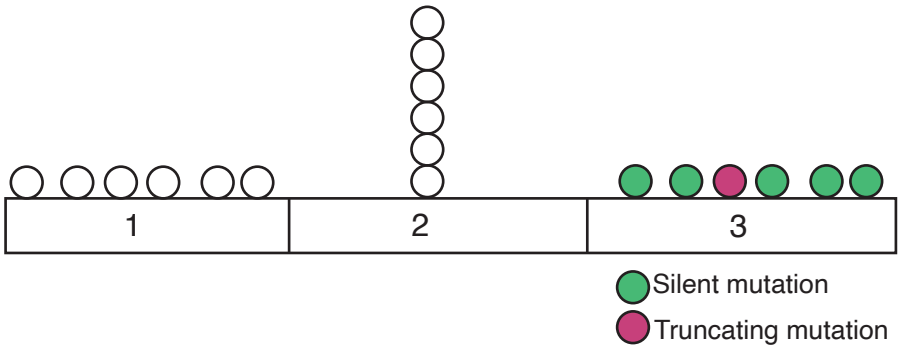


Figure 2.1: Important aspects for driver gene identification. Based on the pysical locations of mutations, the role of a gene in cancer could be identified. While mutations in tumor suppressors tend to be distributed along the gene (1), physically clustered mutations are common for ongocenes (2). In addition, the functional impact of mutations should also be taken into account as many silent mutations with no functional affect might accumulate (3) due to some intrinsic factors (e.g. gene length).

Frequency-based approaches test the significance of a mutated gene based on its frequency in a patient cohort (Dees et al., 2012; Lawrence et al., 2013). In order to assess the significance of observed mutations, these methods predict the background mutation rate integrating different covariates that affect somatic mutation rates such as expression levels or replication timing. One of the champion methods, MutSigCV (Lawrence et al., 2013) integrates a large number of features of the genome and predict the somatic mutation rate of the genome before testing the significance of the observed data.

The second approach in driver gene identification is to assess the functional impacts of accumulated mutations in a cohort (Davoli et al., 2013; Gonzalez-Perez & Lopez-Bigas, 2012). These methods use functional impact scores obtained from tools such as Polyphen or SIFT (Adzhubei et al., 2010; Sim et al., 2012). The aim is to find genes with high functional impact mutations because genes with high functional impact mutations are more likely to be driver genes.

The third approach leverages the observation that oncogenes tend to repeatedly have gain-of-function mutations at fixed positions of the gene (Davoli et al., 2013; Tamborero et al., 2013). By measuring the physical locations of the somatic mutations in a gene, this type of methods looks for genomic positions with highly clustered mutations (Kamburov et al., 2015; Yue et al., 2010).

After many features associated with driver mutations have been investigated, methods that can integrate multiple features start to gain popularity. For example, Davoli et al. (2013) have developed TUSON Explorer which tests for both regional clustering and functional impact of mutations (Davoli et al., 2013). 20/20+ is another combined approach that applies the random forest algorithm and tests for several features including clustering of mutations in certain loci of the genome as well as several gene specific features (e.g. gene length, replication time) (Tokheim et al., 2016). Combining driver identification tools with different approaches can help to find driver genes which may be missed by a single method.

In summary, instead of simply looking at frequency of mutations (Figure 2.1 1), clustering physical locations of mutations might provide extra information

about potential oncogene role of the gene in cancer (Figure 2.1 2). In addition, looking simply at frequency might be misleading when most of these mutations are silent mutations (Figure 2.1 3). In conclusion, integrating methods with different approaches is important for the discovery of unknown drivers.

In this chapter, I explored the mutational landscape of HCC genomes by integrating five big HCC cohorts (n=1349) and aimed to identify novel HCC driver genes using different algorithms. In addition, I have implemented a saturation analysis for driver identification to investigate how sample size might affect the statistical power to detect the drivers with different frequencies. In addition, I conducted association analysis to correlate drivers to clinical features. In the end, I performed pathway analysis in order to understand the functional impact of driver genes.

2.2 Materials and Methods

2.2.1 Datasets used in this chapter

In this chapter, five large cohorts of hepatocellular carcinoma (HCC) genomes were collected for driver identification. These include International Cancer Genome Consortium (ICGC) database (Junjun Zhang et al., 2011) which contains two Japanese cohorts from Riken and National Cancer Center of Japan (n=514) together with the France cohort (n=242). For two other cohorts from US (TCGA, n=373) and Korea (n=231), raw whole exome sequencing data were collected from the GDC database and by collaborating with the authors of Ahn et al. (2014). The ICGC mutation data were downloaded from

ICGC website (https://dcc.icgc.org/releases/release_25).

Somatic mutations were first annotated using Oncotator (Version 1.9.2) using hg19 as the reference build (Ramos et al., 2015). Since the combined dataset contains both the whole exome and whole genome sequencing datasets, samples were uniformized by taking only the coding variants. Hypermutated samples with more than 1000 mutations in coding regions were excluded (n=11).

2.2.2 Somatic mutation calling

For TCGA and Korean cohorts, BAM files were obtained from TCGA database and the authors of Ahn et al. (2014). In order to have a more uniform pipeline, original raw reads were extracted from the BAM files and the same mapping as well as variant calling pipelines were applied. Raw paired end reads were mapped to human reference genome (GRCh37) using Burrows-Wheeler Aligner (BWA) (version 0.7.12) (H. Li & Durbin, 2010). Duplicated reads were removed using sambamba (version 0.6.4) (Tarasov et al., 2015) and base quality recalibration and local realignment were conducted using the Genome Analysis Toolkit (GATK, version 3.1-1). Somatic point mutations were called using Mutect (1.1.7) algorithm by comparing normal and tumor samples (Cibulskis et al., 2013). Strelka (version 1.0.14) was used for indel calling (Saunders et al., 2012).

2.2.3 Curation of known driver genes and pathways

Since driver identification has been conducted in several medium size studies (cohort size of a few hundred or less), the field's latest understanding about HCC drivers were consolidated by compiling a joint list of drivers from eight publications (Ahn et al., 2014; Ally et al., 2017; Chaudhary et al., 2018; Cleary et al., 2013; Fujimoto et al., 2016; Kan et al., 2013; Schulze et al., 2015; Totoki et al., 2014). A total of 88 genes were curated from these eight studies. In order to further annotate potential known drivers, Cancer Gene Census (CGC) genes list from the Catalogue of Somatic Mutations in Cancer (COSMIC) database (GRCh37 / COSMIC v83) was used (Forbes et al., 2017).

To compile the significantly altered pathways in HCC, pathway information from seven different HCC genomic studies were pooled (Ahn et al., 2014; Ally et al., 2017; Dow et al., 2018; Guichard et al., 2012; Kan et al., 2013; Schulze et al., 2015; Totoki et al., 2014). A consensus list of altered pathways was further compiled restricting to those pathways appearing at least twice in these seven studies.

2.2.4 Methods for identifying driver genes

Three different methods were used to identify drivers in this combined cohort. MutSigCV (version 1.41) (Lawrence et al., 2013), 20/20+ (Tokheim et al., 2016) and TUSON Explorer (Davoli et al., 2013) target different aspects of the driver profile. Significantly mutated genes were filtered based on FDR controlled q-values ($q < 0.1$). For both TUSON explorer as well as 20/20+, the

computational algorithm will calculate the probability of being an oncogene (OG) and tumor suppressor gene (TSG) for each driver. For TUSON explorer, Polyphen2 HumVar scores were extracted from the Oncotator output. A final list of driver genes was constructed combining results from all three methods and genes that are mutated in less than 1% of samples were further filtered.

2.2.5 Saturation analysis

To evaluate how the sample size can affect driver identification, a down-sampling strategy was implemented similar to Lawrence et al. (2014). Using a different number of subsets (n=100, 250, 500, 750 as well as 1000) from the total set (n=1349), MutSigCV were used to identify candidate drivers within the subsets (Lawrence et al., 2014). For each sample size, 5 different replicates were carried out. Drivers of different frequencies were further categorized into discrete categories (>20%, 5-20%, 2-5%, <2%) based on their frequencies in the total set.

2.2.6 d_N/d_S calculations

An R package called dndscv (Martincorena et al., 2018) was used by providing driver gene list and mutation file across 1349 patients.

2.2.7 Mutual exclusivity and concurrence analysis

Using the presence and absence matrix, the association between drivers as well as between drivers and clinical phenotypes was tested. One-sided tests were

carried using the Fisher exact test examining the co-occurrence as well as mutual exclusiveness of the variables. Multiple test correction was carried out using the Benjamini-Hochberg method. Statistical associations with a q-value of less than 0.1 were extracted as the candidate list.

2.2.8 Survival analysis of drivers

To identify driver mutations associated with patient survival, a multivariate Cox model is generated by adding in the cohort of patient to the model to adjust for cohort differences. Only drivers with a p-value less than 0.05 were kept. For selected drivers, a univariate analysis is also conducted and univariate log-rank p-value is calculated. Kaplan-Meier survival curves were plotted in R software and both log-rank and Cox model p-values were shown on the plots.

2.2.9 Calculation of cancer cell fraction and mutation timing

Mutation timing was carried out using 579 patients from TCGA and Korean cohorts as required raw WES data required for timing analysis was available for these subset of patients. Sequenza was employed to infer the integer copy numbers using BAM files (Favero et al., 2015). Cancer cell fraction (CCF) of single nucleotide variants was calculated following similar principles to McGranahan et al. (2015). In particular, VAF = purity $\times$ CCF / (CN_{normal} $\times$ (1-purity) + purity $\times$ $CN_{mutation}$) formula was used (equation 1). CN_{normal} is the normal copy number of the loci and $CN_{mutation}$ is the mutation copy number. 2 (diploid) was used for autosomal mutations. For mutations in the X chromosome, 2 was used for female patients and 1 was used for male patients.

Mutation copy numbers were calculated by overlapping mutation location with segments segmented copy number and purity values estimated from the Sequenza (Favero et al., 2015) were used.

Mutations with a total number of reads less than 10, a number of alternative alleles less than 3 and/or allele frequency of less than 0.05 were filtered out. For each mutation, I conducted binomial modeling of observed VAF data using the binomial distribution. A likelihood function is defined by calculating binomial probability using the depth of coverage as the number of trials, the number of mutated alleles as the number of success and VAF was used as a function of CCF based on equation 1. Then, a deviance function was defined as $-2\sum(log - likelihoods)$ using log-likelihood. Finally, the deviance function was optimized between [0,1] interval using optim() function in R to find the CCF value which minimizes the deviance function, meaning the highest value of binomial probability. Timing of the mutation is classified based on cancer cell fraction (CCF). Early mutations were defined as CCF>=0.8 and late mutations were mutations with CCF<0.8.

2.2.10 Pathway analysis

Using the combined driver list (n=62), significantly altered pathways were discovered by combining results from ConsensusPathDB (Kamburov et al., 2009) and g:profiler (Reimand et al., 2016). g:Profiler applies Fisher's Exact test to identify over-represented pathways in a given gene list. The algorithm allows user to order their gene list according to a scoring metric (e.g. mutational frequency). ConsensusPathDB finds pathways enriched in a given gene list

using a hypergeometric test. For g:profiler, driver genes were ordered according to decreasing sample frequency. For both tools, pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa & Goto, 2000) and Reactome (Croft et al., 2014) databases were selected. In addition, pathways from BioCarta (Nishimura, 2001) database were also included in the ConsensusPathDB analysis. Given the output from these two methods, pathways with similar functions were first grouped together. Comparing the literature curated known pathways, this list is further stratified into known pathways as well as novel ones.

2.2.11 Timing pathways

Early and late mutation numbers were summarized for all driver genes within a pathway. These counts were compared with early and late mutation numbers of drivers in all other pathways. 2 x 2 contingency matrix was generated using the mentioned counts and Fisher's Exact test applied to get the odds ratio (OR) and the p-value. As an OR greater than means a higher proportion of late mutations, pathway with an OR greater than 1 were denoted as late pathways and early otherwise.

2.3 Results

2.3.1 Patient cohort characteristics

I collected HCC genomes from five big cohorts comprising of two Japanese (LIRI-JP and LINC-JP from ICGC database, n=514), a France (LIRI-JP from

ICGC database, n=242) (Schulze et al., 2015), a US (LIHC from TCGA, n=373) and a Korean cohort (from Ahn et al. (2014), n=231). I excluded patients with total number of mutations in the coding regions higher than 1000 as these are considered as hypermutated samples which might affect the driver identification result (Table 2). After the elimination of hypermutated samples (n=11), there are 366, 229, 513 and 241 samples from the US, Korea, Japan as well as the French cohort. In total, we have mutation data from 1349 HCC patients together with their clinical information where available.

Table 2: Summary of study cohorts for driver identification

Data source	Project	Cohort	#Samples	Technology	Center	#Hypermutated
TCGA	LIHC	TCGA	373	WES	Broad	7
ICGC	LIRI-JP	LIRI-JP	270	WES/WGS	RIKEN	1
ICGC	LINC-JP	LINC-JP	244	WES/WGS	National Cancer Center	0
ICGC	LICA-FR	French	242	WES	Institut National du Cancer	1
Ahn et al. 2014	-	Korean	231	WES	Asan Bio-Resource Center, Korea	2

The median age of the patients is 63 and is similar across different cohorts. French and part of the TCGA cohort (n=179) are Europeans, the rest of the patients are Asians. Across all cohorts, 74% of all patients are male (n=993). While the proportion of female patients are around 20%-25% across the cohorts, this proportion is 33% in the TCGA cohort and is significantly higher than the rest ($p = 0.007$). Of all disease etiology, 402 patients are non-viral (non B or C), while 392 patients are HBV positive and 300 patients are HCV positive. It is important to note that 88% of all HBV positive cases are Asian patients (p-value = 1.02e-43) which imply a significant etiological difference in HCC. There are no significant differences in the proportion of HCV positive cases in

different ethnic groups (p-value = 0.17, See Table 3).

Table 3: Summary of clinical characteristics in the combined HCC cohort

Features	French	Korean	LIRI-JP	LINC-JP	TCGA	Total
Age						
Median (range)	65 (17-90)	55 (26-80)	69 (31-89)	66 (23-85)	61 (16-90)	63 (16-90)
Gender						
Male	190	174	201	182	246	993
Female	47	55	68	62	120	352
Missing	4	0	0	0	0	4
Tumor Stage						
Early	127	229	163	85	254	858
Late	55	0	106	110	89	360
Missing	59	0	0	49	23	131
Etiology						
HBV	31	167	66	–	99	363
HCV	59	21	145	–	48	273
HBV+HCV	2	0	4	–	7	13
NBNC	143	41	53	–	193	430
Missing	6	0	1	244	19	270
Ethnicity						
European	241	0	0	0	179	420
Asian	0	229	269	244	159	901
Missing	0	0	0	0	28	28

2.3.2 Mutational landscape across cohorts

Combining mutation data across 1349 tumors, the final dataset consists of 129,292 SNPs and 7,242 indels. The majority of the mutations (62.5%) are missense and silent mutations are constituting the 23.8% of the total SNVs (Figure 2.2 A). Overall, the mutation types are similar across cohorts (Figure 2.2 B). While the mutation burden is highly variable across patients,

the median mutation rate is also similar across cohorts (median is around 2.87/Mb) (Figure 2.2 C-D).

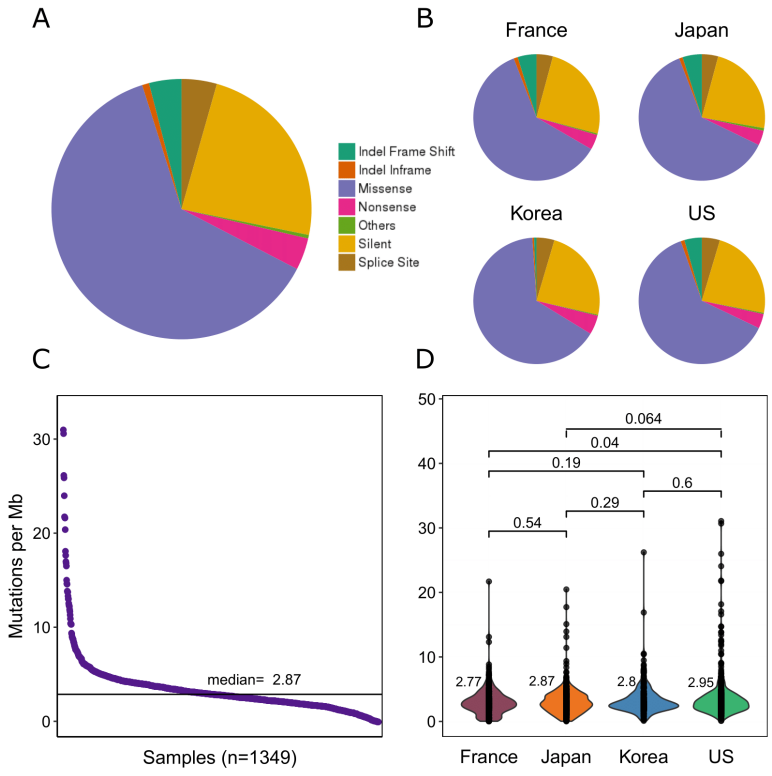


Figure 2.2: Summary of mutation burden. (A) Overall distribution of mutation types across all patients. (B) Distribution of mutation types in different cohorts. (C) Distribution of mutation rates (i.e. number of mutations per megabase across patients.) (D) Mutation rate distributions in different cohorts. Median mutation rates are shown for each cohort. P-values from the Wilcoxon signed-rank test across cohorts are listed for each comparison.

2.3.3 Driver genes in hepatocellular carcinoma

Even though sample sizes are quite decent in many previous studies, they each focus on a specific ethnic cohort with different etiological subtypes. A large

sample size allows us to aggregate mutational information from multiple cohorts and increase the power of identifying novel drivers. Since driver genes tend to have several distinct characteristics, I employed MutSigCV (Lawrence et al., 2013) (a method based on mutation rate), TUSON Explorer (Davoli et al., 2013) (a method based on clustering of functionally important mutations) as well as 20/20+ (Tokheim et al., 2016) (a machine learning approach unifying multiple features of somatic mutations) to identify drivers in this combined cohort.

In total, 45 genes from MutSigCV (Lawrence et al., 2013) , 16 genes from TUSON Explorer (Davoli et al., 2013) and 39 genes from 20/20+ results are identified (Tokheim et al., 2016). A large proportion of these drivers are overlapping between different methods and 62 driver genes are identified combining results from different methods (Figure 2.3, Table S1, Table S2, Table S3, Table S5). Across the 62 drivers, 13 genes including *TP53*, *CTNNB1*, *DYRK1A* and *NFE2L2* are identified by all three methods and all of them (Figure 2.4 A) were previously reported as driver gene except for *DYRK1A*. Litovchick (2011) reported that *DYRK1A* phosphorylates a component of DREAM complex (LIN52) which induce quiescence in cells. A possible mechanism for driving HCC can be escaping from quiescence for aberrant cells because of inactivating mutations in *DYRK1A* gene.

In order to compare my new driver list against published results from the research community, I conducted a literature review of several large cohort studies (n=8) to compile reported drivers in the field (Table S4). In total, I curated 88 known drivers from these eight studies (Ahn et al., 2014; Ally

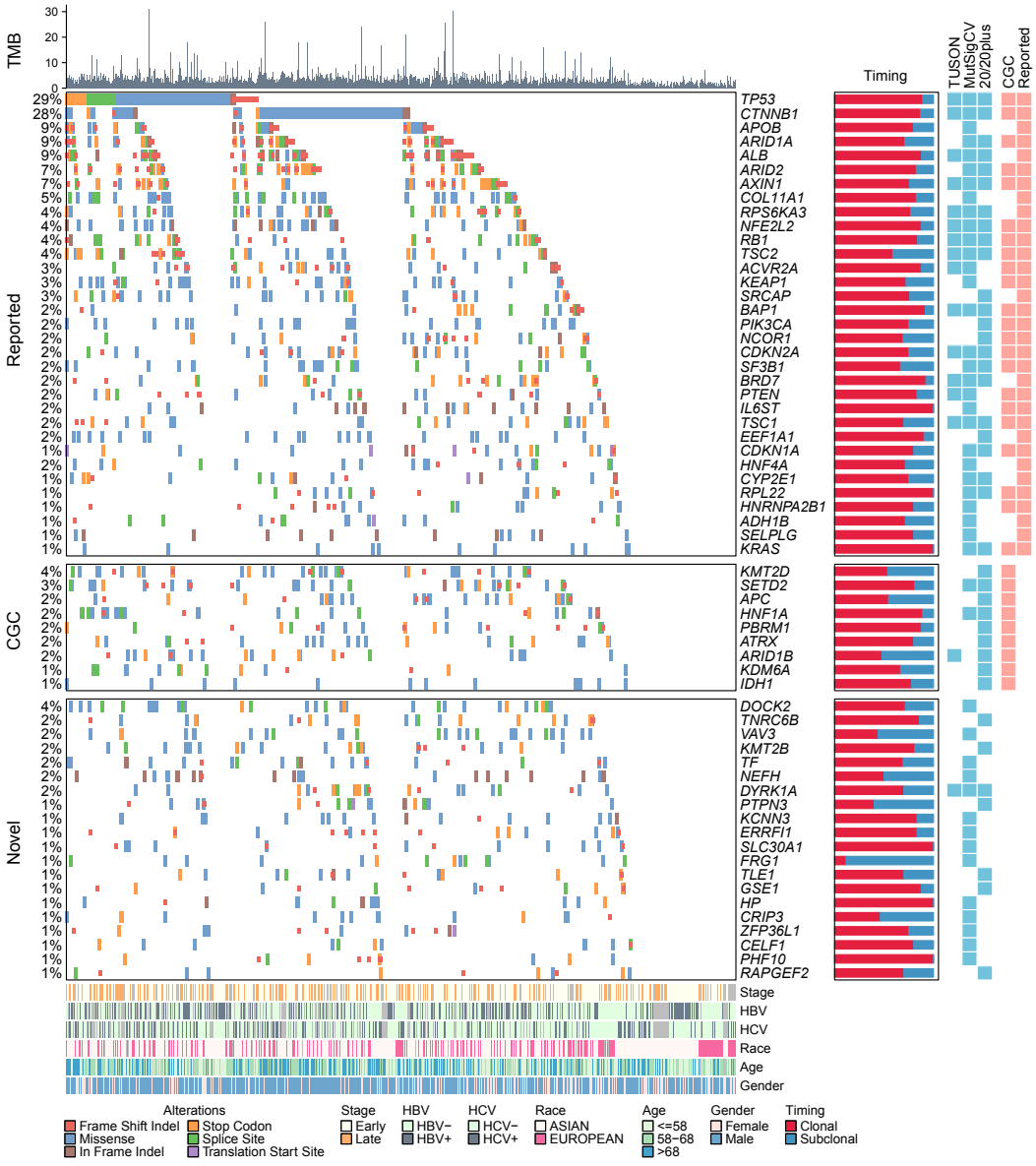


Figure 2.3: The landscape of HCC drivers. Each column represents a patient. In each row, alterations (represented by different colors) and mutation frequencies for each driver is given. The barplot on the right side shows the number of SNVs in the gene across patients. The proportion barplot on the right side indicates proportions of early (red) and late (blue) mutation proportions in the driver across the cohort. Heatmap at the right side indicates whether the driver gene is detected by different methods (TUSON Explorer, MutSigCV and 20/20+) or whether the gene was reported previously by other studies or in the cancer gene census list (Reported or CGC). Barplot on the top indicates the mutation burden of patients. Clinical phenotypes of patients are shown at the bottom of the figure.

et al., 2017; Chaudhary et al., 2018; Cleary et al., 2013; Fujimoto et al., 2016; Kan et al., 2013; Schulze et al., 2015; Totoki et al., 2014). While the majority of these previously reported drivers (76%) were identified by only one of these eight studies (Figure 2.5), as few as 21 genes (24%) were identified by two or more studies (Figure 2.5). Moreover, only two classical driver genes, *TP53* and *CTNNB1*, were identified by all studies. Thus, previous HCC driver identification studies were highly discordant. I next compared my identified driver list (n=62) with the literature drivers (n=88). 33 driver genes overlap between literature driver genes and my new list. Among the 29 novel drivers, 9 of them appear in the Cancer Gene Census (CGC) while 20 of them are novel. A good concordance (53%) with previous studies implies that my integrative method is robust (Figure 2.4 B). Among CGC driver genes (n=9), *ATRX*, was found to be associated with alternative lengthening of telomerase (ALT) which leads to telomere synthesis without telomerase in some sarcomas (Lawlor et al., 2019). Another important driver gene *PBRM1* is a chromatin remodeling gene and has also been found to be frequently mutated in high-grade clear cell renal carcinoma (Gerlinger et al., 2014; Wang et al., 2020). Among completely novel drivers, *FRG1* is a Facioscapulohumeral muscular dystrophy (FSHD) related gene and the increased expression of *FRG1* can increase cell migration and invasion in cancer cell lines (Tiwari et al., 2019, 2017).

In addition to the identification of the driver genes, TUSON Explorer and 20/20+ calculate the probability of oncogene (OG) or tumor suppressor gene (TSG) for each driver. In total, 30 candidate driver genes are classified as TSG and 9 candidate drivers are classified as OG (Table S6). This list includes known

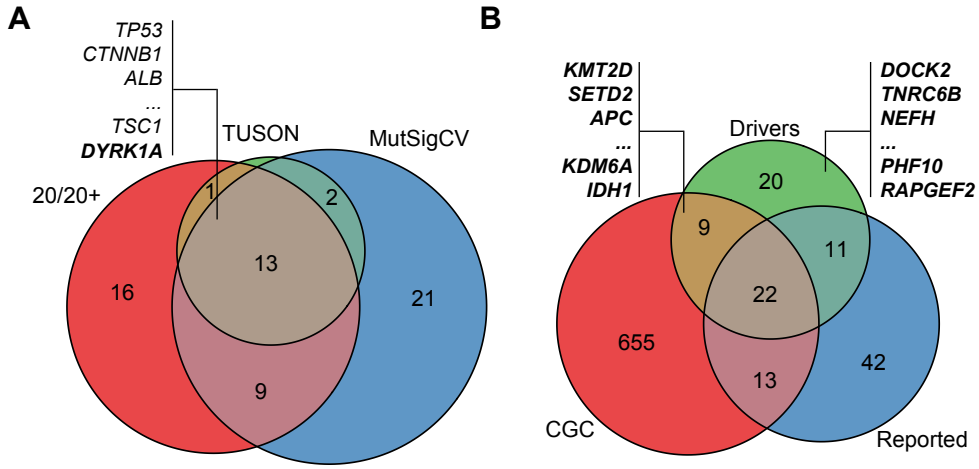


Figure 2.4: Overlap of different driver lists (A) Venn diagram of overlapping and unique drivers identified by individual algorithms. (B) Venn diagram of overlapping and unique drivers identified in this thesis, previous studies (Reported) and in the Cancer Gene Census (CGC) database. Novel genes are written in bold font.

oncogenes such as *CTNNB1*, *PI3KCA*, *NFE2L2*, *KRAS*, *EEF1A1* and *SF3B1* as well as new HCC oncogenes such as *IDH1*, *KCNN3* and *CRIP3* are novel HCC drivers. A known feature of oncogenes is a clustering of mutations on certain regions of the gene. While some of the identified oncogenes have highly frequent mutations at single positions (e.g. *CTNNB1* n=49), some only have a few mutations in a single position potentially due to low mutation rate across the cohort (e.g. *KCNN3*, n=5) (Figure 2.6).

2.3.4 Saturation analysis of driver identification

Since the statistical power of driver identification methods tends to increase with the sample size, I thus performed subsampling analysis asking whether the number of drivers has reached saturation. By down-sampling the cohort

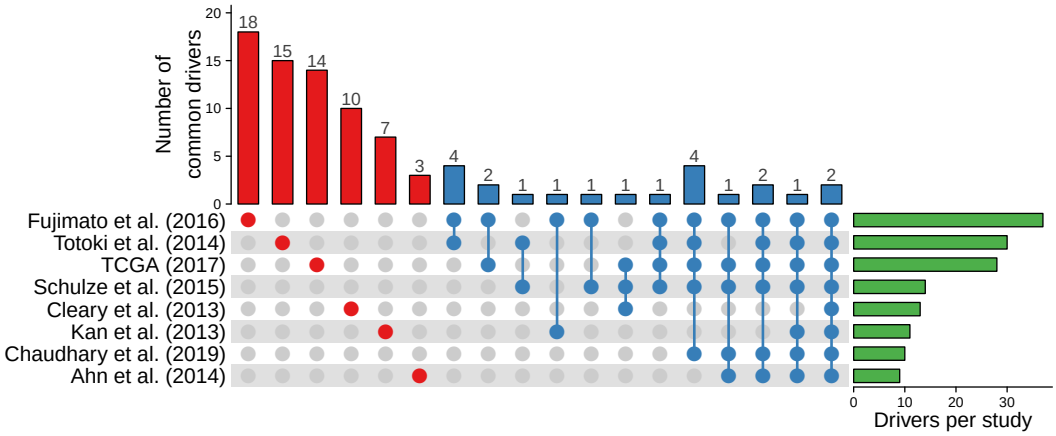


Figure 2.5: Large scale HCC driver identification studies. Right bars indicate the number of identified driver gene in each study. Top bars indicate the number of genes shared by all possible combination of studies. Blue bars denote the number of genes identified by at least two studies and red bars indicate the number of genes identified by only one of the studies. While the majority of driver genes are identified by only one study, only two driver genes are identified across all studies.

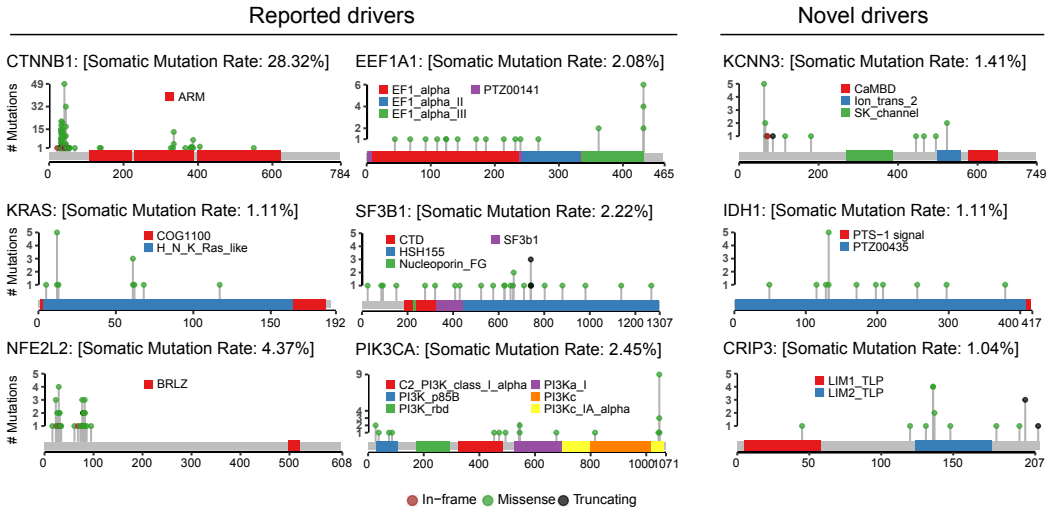


Figure 2.6: The lollipop plots for the identified oncogenes. Domains are shown as colored segments. Mutations with different colors are shown as lollipops. X-axis is amino acid positions and y-axis is the number of mutations in the cohort (n=1349).

to various subsets, driver identification was implemented iteratively using MutSigCV. Interestingly, the number of identified driver genes increased linearly with the sample size and has not reached saturation (Figure 2.7 A).

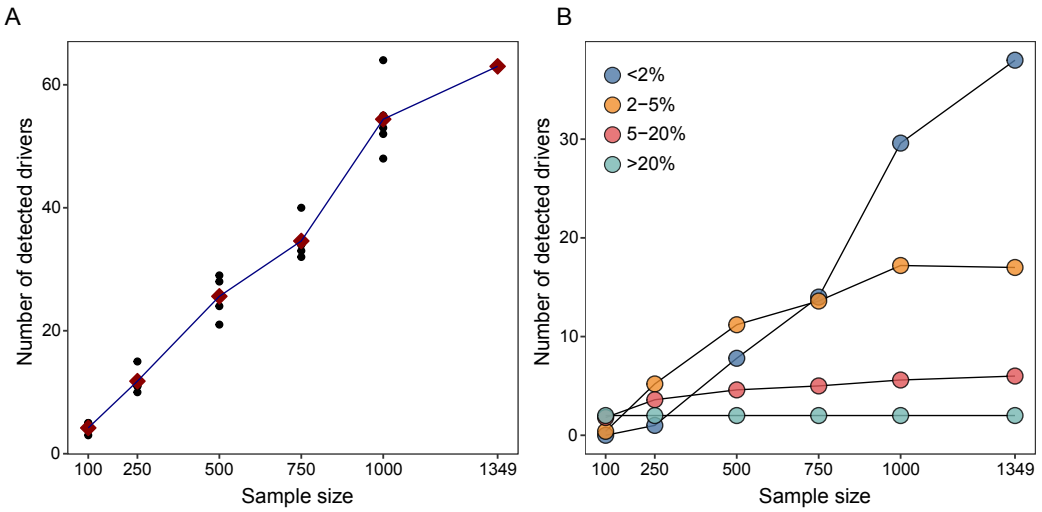


Figure 2.7: Number of drivers detected with different sample sizes. (A) The relationship between the number of identified driver genes and the sample size. Lines are connecting the median number of detected drivers for each subsample (5 replicates per sample size). The number of identified drivers linearly increase with the increasing sample size. (B) Drivers in different frequency categories across samples. While a number of high frequency drivers reaches saturation rapidly, lower frequency drivers are much harder to detect and have not reached saturation.

When breaking down into categories according to their prevalence, I found that drivers with a frequency above 5% are discovered at the sample size of a few hundred (Figure 2.7 B). Drivers with lower prevalence (e.g. between 2-5%) require a larger sample size of around 1,000 to reach saturation (e.g. at least 20 occurrences or higher). For extremely rare drivers with an overall frequency of less than 2%, a sample size much larger than the current cohort will be needed and the number of identified drivers is still rapidly increasing with the sample size. This might explain the observation that a large proportion (14.6%) of the

HCC patients have no drivers identified. Overall, these observations suggest that there are still many extremely rare drivers to be discovered in HCC.

2.3.5 Clinical associations of driver genes

In order to explore the correlation between drivers and patient clinical phenotypes, I carried out association analysis between driver status and clinical features. Not surprisingly, many drivers tend to occur in patients with the higher mutational burden (above median) and higher age (above 63) (Figure 2.8). When comparing drivers across races, only *TP53* (higher in Asians) and *HNF1A* (higher in Europeans) tend to segregate differently between Asian and European cohorts. This indicates that heterogeneity in driver prevalence across different patient cohorts is quite minor. When comparing HCC from different etiologies, HBV positive tumors tend to have the largest number of co-occurring drivers. The classical driver *TP53* is found to be associated with HBV infection. In addition, *ACVR2A* and *CTNNB1* genes tend to co-occur with HBV negative patients. On the other hand, *CTNNB1* mutations are enriched in HCV positive patient group. Other clinical features such as gender, tumor stage and microvascular invasion (MVI) only weakly correlated with a handful of drivers. It is important to note that, a chromatin remodeling gene *ARID2* mutations significantly higher in late stage tumors. In addition, cohort adjusted multivariate survival analysis (see Methods) revealed driver genes that stratify patients such as *TP53*, *FRG1* and *COL11A1*. All of the stratifying drivers causes poor overall survival when mutated (Figure 2.9). Association of *TP53* and *RB1* with poor survival were identified before (Ahn

et al., 2014).

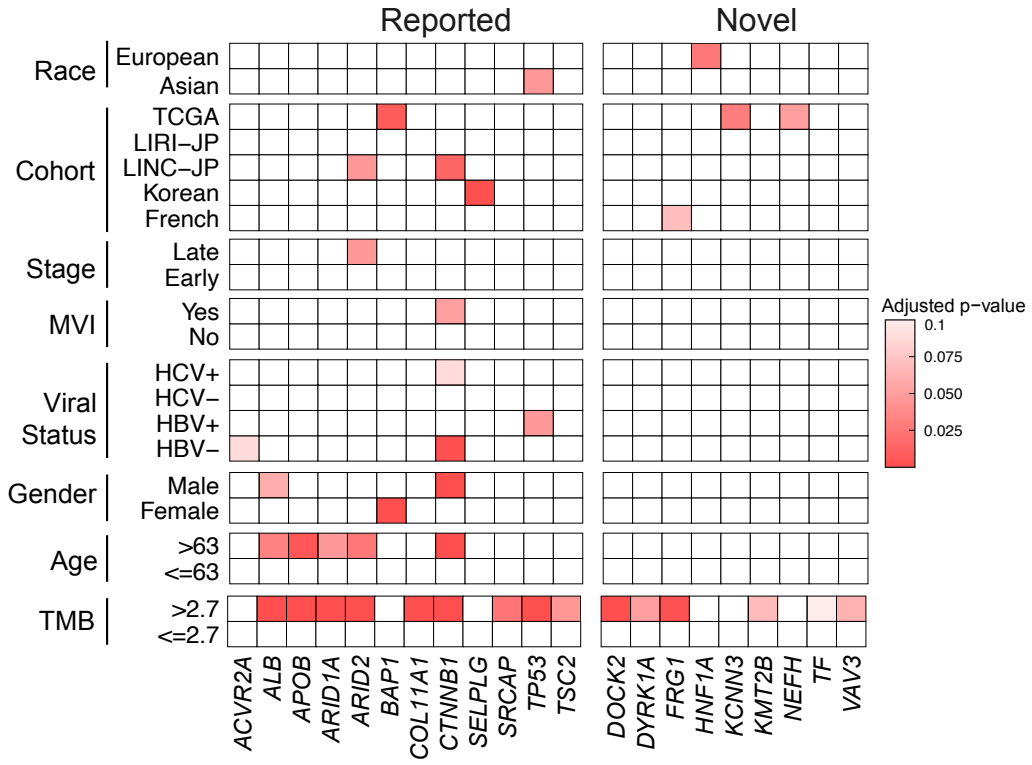


Figure 2.8: Co-occurrence of drivers with clinical phenotypes. One sided Fisher's Exact test (greater tail) was applied to test co-occurrence of the feature with driver genes. Color scale represents the adjusted p-value after Benjamini-Hochberg multiple test correction.

2.3.6 Mutual exclusivity and co-occurrence of driver genes

Since carcinogenesis depends on gaining a full collection of hallmarks associated with cancer, drivers with similar functional roles from the same pathway will tend to occur exclusively with each other. On the other hand, genetic changes with complementary functional roles might act synergistically and tend to co-occur in the same patient. By testing the association between the driver

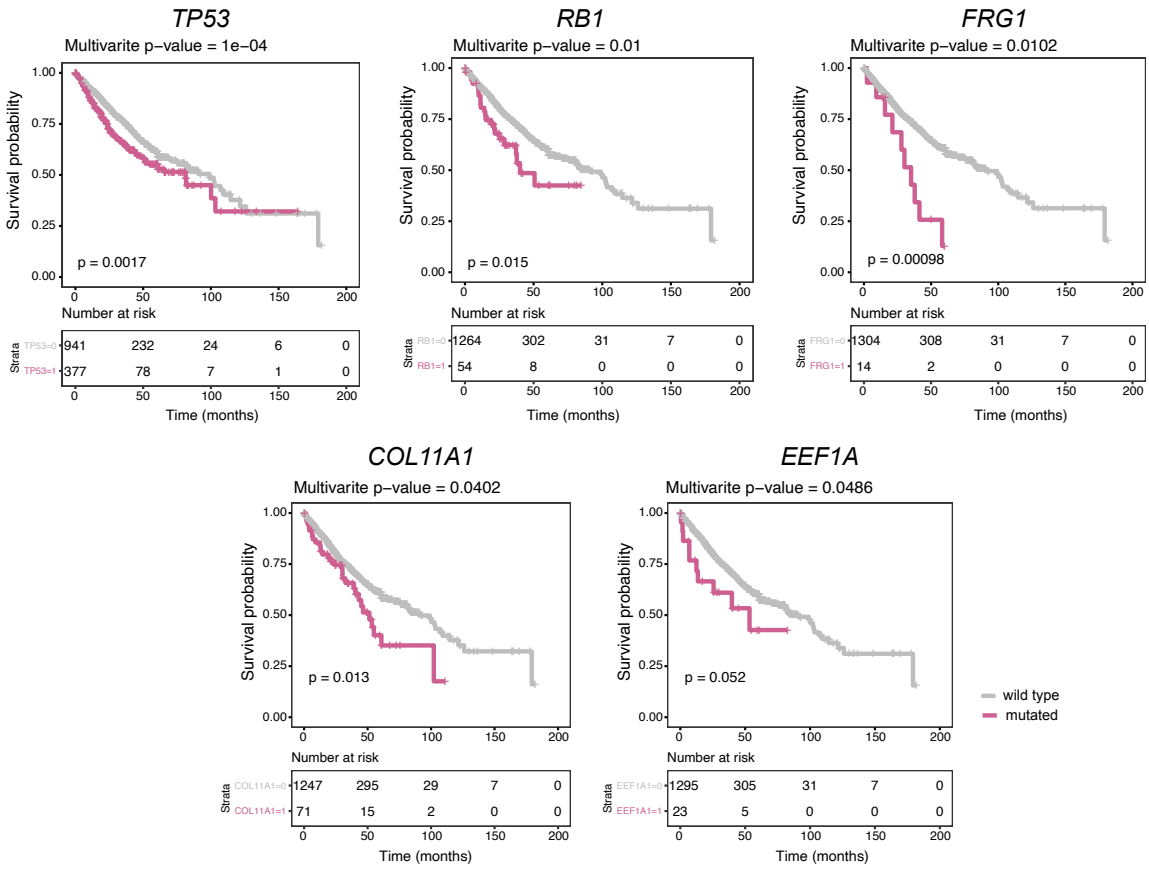


Figure 2.9: Survival curves of driver genes that stratify patients. Driver genes are selected by generating a multivariate Cox model with the cohort variable to adjust for cohort differences. Univariate log-rank p-values for selected drivers are also shown inside Kaplan-Meier plots.

mutations, I identified 11 pairs of co-occurrence relationship and 2 pairs of mutual exclusivities after multiple testing correction (Figure 2.10, Table 4). For example, *CTNNB1* and *AXIN1*, activators of the WNT pathway, tend to occur mutually exclusively across patients. On the contrary, multiple driver pairs tend to co-occur in multiple patients. For instance, *CTNNB1* and *NFE2L2*, oncogenes of the WNT and the oxidative stress pathways tend to co-occur in many patients, suggesting a synergistic effect between these pathways in HCC. The distribution of mutations and mutation types are shown in Figure 2.11.

Table 4: Mutual exclusivity and concurrence between drivers

Gene 1	Gene 2	pvalue_less	pvalue_greater	q-value less	q-value greater	status
ARID1A	NFE2L2	0.9998524	0.0005456	1.0000000	0.0937870	Concurrent
ARID2	CTNNB1	0.9999971	0.0000078	1.0000000	0.0039934	Concurrent
AXIN1	RPS6KA3	0.9999976	0.0000131	1.0000000	0.0049542	Concurrent
BAP1	IDH1	0.9999811	0.0003781	1.0000000	0.0802669	Concurrent
CRIP3	NFE2L2	0.9999876	0.0001997	1.0000000	0.0539536	Concurrent
CRIP3	SRCAP	0.9999707	0.0005365	1.0000000	0.0937870	Concurrent
CTNNB1	NFE2L2	0.9999997	0.0000014	1.0000000	0.0012795	Concurrent
CTNNB1	PHF10	0.9999882	0.0001279	1.0000000	0.0403219	Concurrent
KDM6A	TP53	0.9999449	0.0003820	1.0000000	0.0802669	Concurrent
KDM6A	TSC2	1.0000000	0.0000001	1.0000000	0.0002118	Concurrent
TP53	TSC2	0.9999976	0.0000084	1.0000000	0.0039934	Concurrent
AXIN1	CTNNB1	0.0000025	0.9999995	0.0046858	1.0000000	Exclusive
CTNNB1	TP53	0.0000060	0.9999969	0.0056697	1.0000000	Exclusive

2.3.7 Pathway analysis of driver genes

In order to understand the commonly perturbed pathways, I performed statistical analysis of 62 driver genes using g:Profiler (based on Fisher's Exact test on a rank gene list) and ConsensusPathDB (based on a hypergeometric

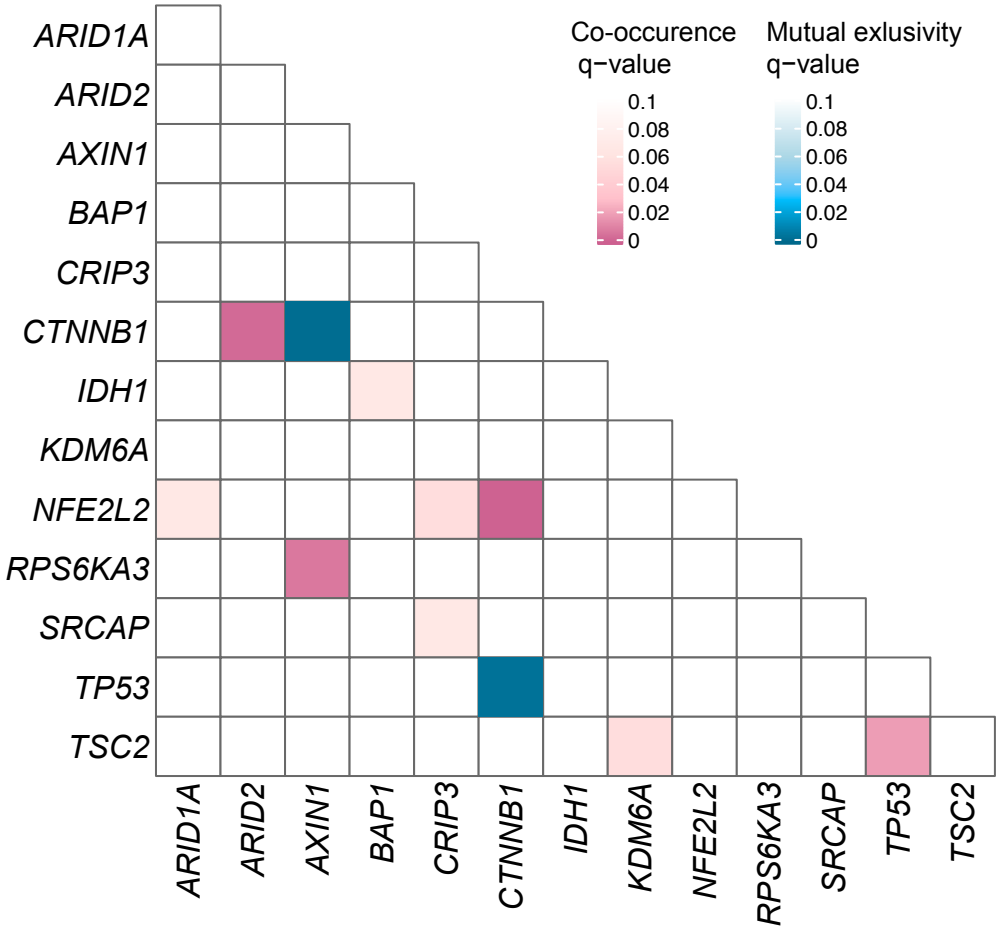


Figure 2.10: Mutual exclusivity and co-occurrence status of drivers. Left tailed Fisher's Exact test was applied to test mutual exclusivity and the right tailed Fisher's Exact test was applied to test co-occurrence of drivers. Benjamini-Hochberg multiple correction test was applied and pairs with a q-value < 0.1 are kept. The q-value each significant mutually exlclusive or co-occurrent pair are shown as gradients of pink (co-occurring) and blue (mutually exclusive).

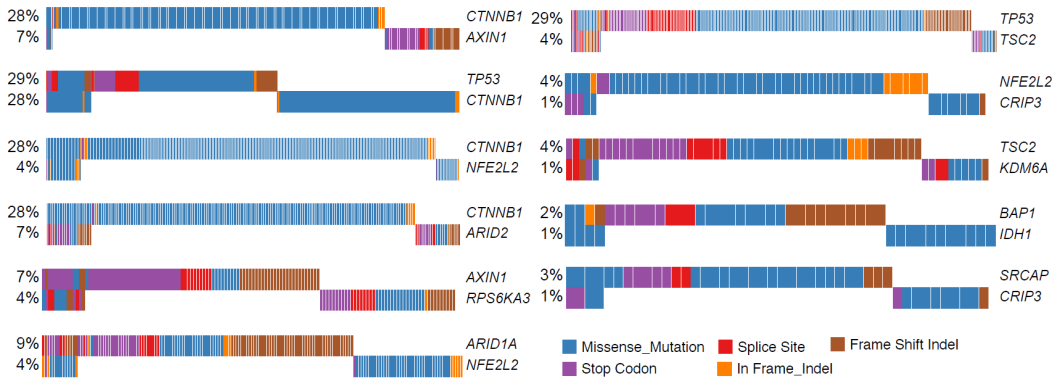


Figure 2.11: Mutually exclusive and co-occurring mutations across the cohort. All 11 pairs of mutually exclusive and co-occuring genes are shown together with their mutation types in different colors. Mutation percentage of each gene is shown at the left side of each row.

test). After combining pathways with similar names, 13 pathways are identified using the g:Profiler method and 24 pathways are identified from ConsensusPathDB method (Table 5, Figure 2.12). In order to compare this list of pathways against previously published results, I compiled 17 significantly perturbed HCC pathways from seven different studies (Ahn et al., 2014; Ally et al., 2017; Dow et al., 2018; Guichard et al., 2012; Kan et al., 2013; Schulze et al., 2015; Totoki et al., 2014).

Table 5: Pathways identified using CPDB and g:profiler and driver genes

Literature Pathways	Novel Genes	CGC Genes (Novel for HCC)	g:Profiler	CPDB	isNovel
WNT	TLE1, TNRC6B	HNF1A, KMT2D, APC	Y	Y	N
TP53/Cell cycle	TNRC6B, DYRK1A	___	Y	Y	N
Chromatin remodeling	KMT2B	ARID1B, PBRM1, KMT2D, SETD2, KDM6A	Y	Y	N
PI3K/AKT/mTOR	TNRC6B	___	Y	Y	N
Apoptosis	___	APC	Y	Y	N
MAPK	RAPGEF2	___	N	Y	N
Adhesion/ECM	VAV3, RAPGEF2	___	N	Y	N
Stress	TNRC6B	___	Y	Y	N
JAK/STAT	___	___	N	Y	N
DNA Damage/Repair	___	___	N	Y	N
NOTCH	TLE1, TNRC6B	___	N	Y	N
TERT/TELOMERASE	___	___	N	Y	N
RTK	TNRC6B, RAPGEF2, PTPN3, VAV3	___	Y	Y	N
Immune	VAV3, DOCK2, TNRC6B, RAPGEF2, HP	IDH1	Y	Y	N
Hypoxia	TF	___	N	Y	N
Virual infection	___	APC	Y	Y	N
Scenescence	TNRC6B; RAPGEF2	___	Y	Y	N
Stemness	TNRC6B, ZFP36L1	APC, HNF1A	Y	Y	Y
Mineral absorption	TF, SLC30A1	___	N	Y	Y
Hemostasis	TF, VAV3, DOCK2	___	N	Y	Y
Actin cytoskeleton	VAV3	APC	N	Y	Y
Thyroid hormone	___	___	Y	Y	Y
Carbon metabolism	___	IDH1	Y	Y	Y

Among the 17 known pathways (Figure 2.12 top left), 13 novel drivers are found to be in these pathways. The discovery of novel genes in known HCC pathways may help to further elucidate the mechanisms of already known pathways. For example, several novel drivers such as *TLE1*, *TNRC6B*, *APC*, *KMT2D* and *HNF1A* are found in the classical WNT pathway (*APC*, *KMT2D* and *HNF1A* are CGC genes but are not reported as driver for HCC) (Figure 2.12 top right).

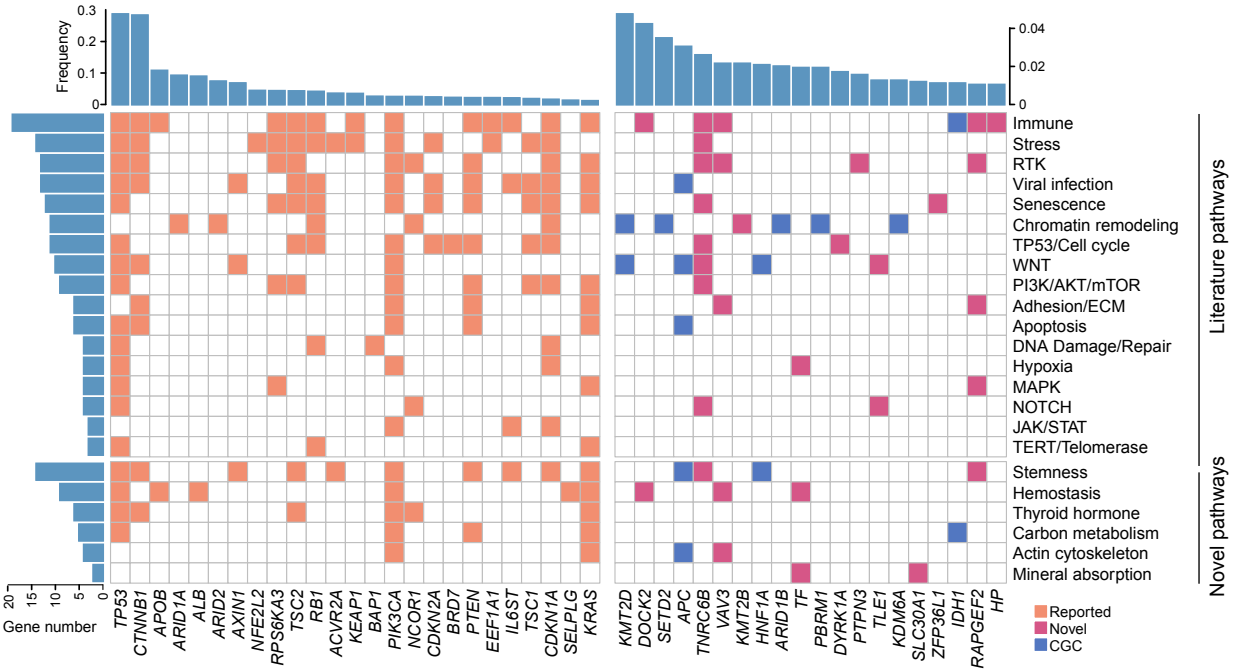


Figure 2.12: Driver genes and pathways in which they function. Both literature and novel pathways are shown. The upper panel bar chart shows the mutational frequency of the corresponding driver gene. The left panel bar chart shows the number of genes assigned to corresponding pathways. Heatmaps on the left show pathways for previously reported driver genes and heatmaps on the right show pathways for novel drivers. Genes from CGC are shown in the same color (orange) on both sides.

In addition to literature known pathways, 6 novel pathways are identified with genes from both reported and new drivers. 14 of previously known HCC drivers

are assigned to the novel pathways (e.g. *CDKN1A* and *ALB*) (Figure 2.12 bottom left). Finally, 9 novel genes (3 of them are in CGC genes) are found to function in identified novel pathways. For example, *SLC30A1* and *TF* novel drivers are allocated to the mineral absorption pathway (Figure 2.12 bottom right).

2.3.8 Clonal status of HCC drivers and pathways

For a given driver mutation presented in the tumor, an important question often raised in the field is the clonal status of the driver. Mutations arising early in tumorigenesis will be present in all cancer cells (i.e. clonal). Given the allele frequency as well as local copy number, we can estimate the proportion of cancer cells carrying this mutation. Since clonal mutations tend to arise early in tumorigenesis, I ask the question what are the drivers that tend to arise early in tumorigenesis. When I plot the proportion of the times a given driver is clonal, there are multiple drivers such as *IL6ST* and *KRAS* who are clonal across all cases. Known drivers such as *TP53* and *CTNNB1* are often clonal in a very high fraction of tumors. This suggests that there are a large number of early HCC drivers in tumorigenesis (Figure 2.13 A).

Interestingly, among identified drivers, novel ones have a higher fraction of late mutations (Figure 2.13 B, C). Moreover, driver genes with lower frequencies tend to occur late in the history of tumorigenesis compared to frequent drivers (Figure 2.13 D). This indicates the importance of the big sample size to identify mutations that are rare and driving tumor progression rather than initiation.

Since many of these drivers fall into a set of known and novel pathways, I then

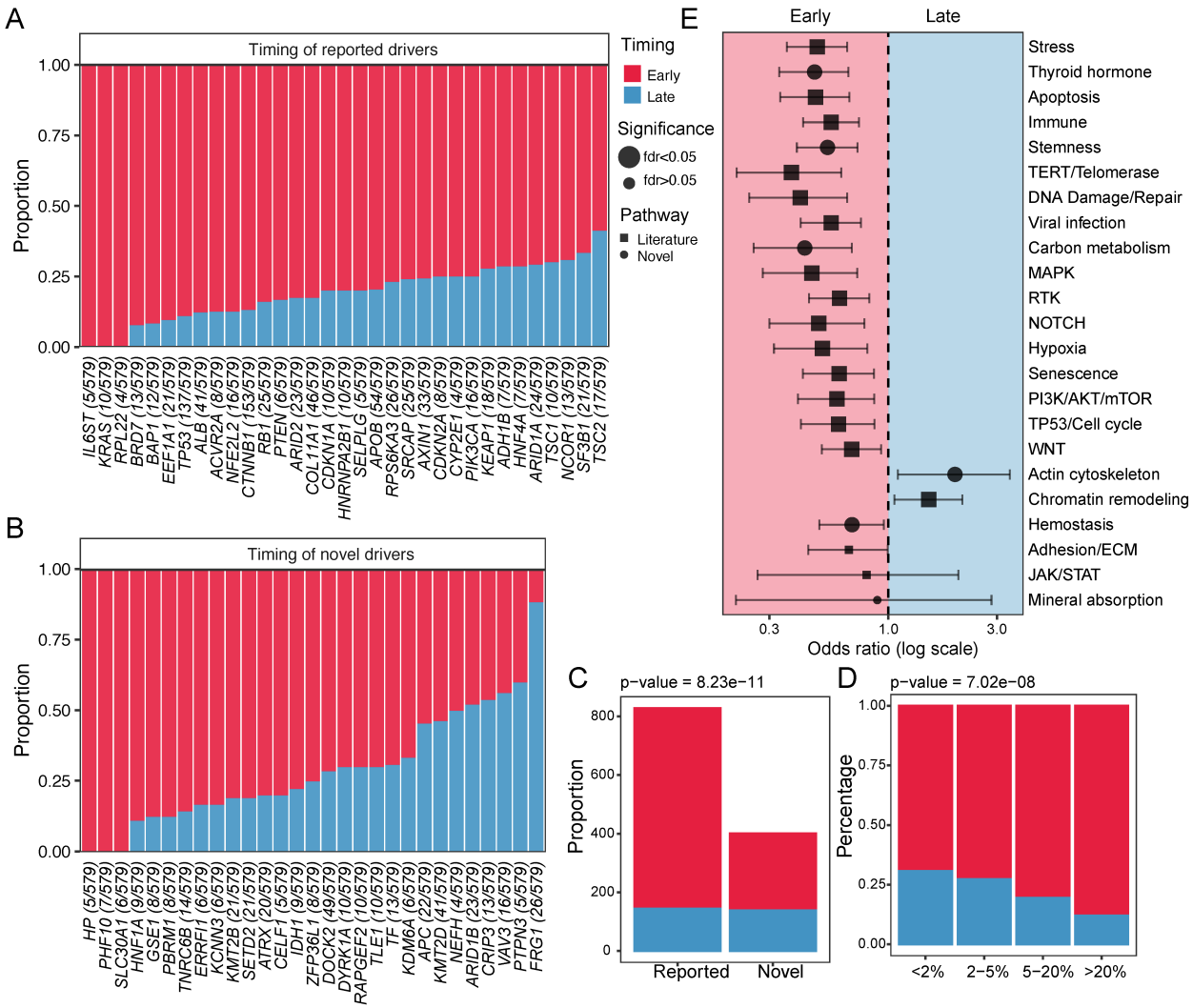


Figure 2.13: Clonal status of HCC drivers and pathways (A) Subclonal (blue) and clonal (red) fractions of mutations in reported drivers across all patients. The number of total mutations in each gene are indicated in brackets. For example, all 5 mutations in IL6ST gene are early. (B) Subclonal (blue) and clonal (red) fractions of mutations in novel drivers across all patients. Overall, a higher proportion of late mutations are observed in novel drivers. For example, nearly all FRG1 gene mutations (n=26) are late. (C) Comparison of early and late mutation proportions between reported and novel drivers. Fisher's Exact test was applied to compare counts of early and late mutations in (D) Comparison of early and late mutation proportions between different driver frequency categories. (E) Estimation of pathway timing.

ask the question of whether certain pathways tend to be perturbed early in tumorigenesis and have predominantly clonal mutations. By comparing the proportions of early and late mutation within each pathway, I found that most of the pathways such as stemness and TERT often occur very early, while chromatin remodeling and actin cytoskeleton pathways tend to be late events, possibly driving further tumor progression (Figure 2.13 E). By "dating" the molecular events along with the history of tumorigenesis, clonality analysis provides a unique angle understanding the molecular cascades in HCC.

2.3.9 Positive selection in drivers

Driver genes that are positively selected in the history of tumorigenesis are known to have a higher rate of nonsynonymous mutations than expected by chance. In order to understand the role of positive selection in the history of tumorigenesis for these driver genes, I calculated somatic non-synonymous to synonymous mutations (d_N/d_S) for all the driver genes using a method from a study (Martincorena et al., 2018). It is found that 56 out of 62 drivers have a significant d_N/d_S value after multiple test correction (q value<0.1, Table S7, Figure 2.14 A). In the method I used (Martincorena et al., 2018), "nonsynonymous" changes are categorized into truncating and missense mutations. When we look at d_N/d_S values calculated for truncating and missense mutations, we can make inference about the role of a gene in cancer as a tumor suppressor (TSG) or an oncogene (OG) as TGSs tend to lose their function via truncating mutations. Hence, a significant truncating d_N/d_S value might be an indicator of TGS while a gene with only missense significant

d_N/d_S value indicates oncogenic potential.

Oncogenes such as *CTNNB1* have missense mutational hotspots and they usually have significant missense d_N/d_S ratios while tumor suppressor genes such as *AXIN1* mostly have truncating mutations and significant truncating d_N/d_S ratios. A well known tumor suppressor gene *TP53* has significance in both truncating and missense d_N/d_S values because inactivation for *TP53* can be achieved by truncation and changes in protein p53 by mutational hotspots such as R249S (Olivier et al., 2010). Thus, clustered missense mutations might also be inactivating and if a gene has both truncating and missense significance in d_N/d_S values, it is more likely to be a TSG.

It is interesting to observe that most of the driver genes have a significant d_N/d_S ratio based on truncated mutations (45%) indicating putative tumor suppressor roles. Only 9 of them exclusively have significant missense d_N/d_S ratios which imply that those drivers are likely to be oncogenes. When we plot truncated vs missense d_N/d_S values we can see the clear separation of tumor suppressor genes and oncogenes (Figure 2.14 B). Interestingly, novel drivers with a significant truncating or both truncating and missense d_N/d_S value have a higher proportion of late mutations (Figure 2.14C-D). Taken together, through an integrative analysis, I have identified a large number of novel drivers that tend to enrich for tumor suppressors and often occur late during tumor progression (Figure 2.14C-D).

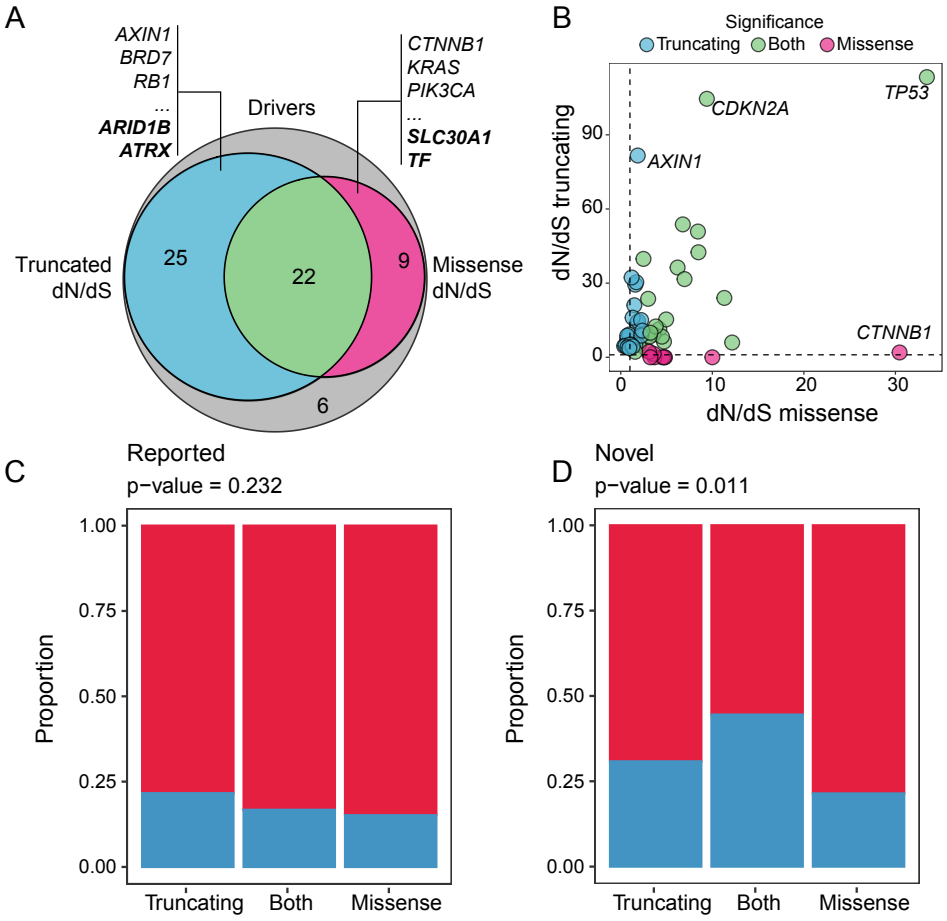


Figure 2.14: Summary of dN/dS analysis (A) Venn diagram of drivers with significant truncating or missense dN/dS value. (B) Truncating versus missense dN/dS values of driver genes. Comparison of early and late mutation proportions across genes with significant truncating, missense or both dN/dS values for (C) reported drivers and (D) novel drivers. Fisher's Exact test was applied for the comparison of proportions.

2.4 Discussions

2.4.1 Selection criteria for analysis datasets

In this chapter, main analyses are driver identification and timing analysis of driver mutations. For driver identification, needed data is position and base change for mutations while copy number data is also needed for timing analysis. For driver identification, I collected somatic mutation data from five different cohorts (LIRI-JP, LINC-JP, LICA-FR, TCGA, Korean)) and these cohorts comprises all open access HCC cohorts with available mutation data to my knowledge. As copy number data across the genome was not usually publicly available, I needed raw sequencing data for timing analysis of drivers. I was only able to use TCGA and Korean cohorts for timing analysis as these two were only datasets which I was able to collect the raw DNA sequencing data and call copy numbers (via our lab access to TCGA database and data request from the authors of the Korean cohort (Ahn et al. 2014) paper).

2.4.2 Rationales for bioinformatics tool selection

Driver gene identification: Since driver genes tend to have several distinct characteristics such as being frequently mutated across the cohort, having mutations with important functional impact, I aimed to use multiple driver identification algorithms to capture drivers with different features. While MutSigCV (Lawrence et al., 2013) is a method based on comparison of observed mutations rate to the backgorund mutation rate, TUSON Explorer (Davoli et al., 2013) is based on clustering of functionally important mutations.

In addition, 20/20+ algorithm is a machine learning method which integrated several features (Tokheim et al., 2016) and I employed this algorithm to capture complex features of drivers.

Timing analysis: McGranahan et al. (2015) calculated the timing of driver genes across many cancer types using TCGA data. However, this study does not include HCC patients. Implemented method in McGranahan et al. (2015) is accessible as an R package (EstimateClonality). However, as this method does not calculate timing of mutations in genes located in sex chromosomes and there are sex chromosome genes in my driver list (e.g. *ATRX* and *KDM6A*), I re-implemented this method by adding this feature. For common mutations, original method and my modified method showed very high concordance (Sprearman's rho=0.956).

2.4.3 Comparison of HCC mutation burden and drivers with other cancer types

Different cancer types arise from different mutational processes and it is normal to expect different mutational landscape across cancers from different tissues due to different dependencies on different pathways. In this chapter, with the largest combined HCC dataset (n=1349), a median tumor mutation burden (TMB) of 2.7 Mutations per megabase is observed. To compare this TMB with other cancer types, I downloaded data from a pan-cancer which reported TMB across cancer types and did not include HCC (Lawrence et al., 2013). Figure 2.15 summarizes the distributions of TMB across different cancer types. Cancers which are known to be associated with exposure to known carcinogens

such as melanoma (e.g. UV lights) and lung cancer (e.g. smoking) rank first in TMB ranking (Figure 2.15). On the other hand, cancer types which frequently emerge in children such as rhabdoid tumors and medullablastoma are the ones with the lowest TMB. Interestingly, HCC shows intermediate levels of TMB compared to other cancer types. Driver genes are also very different for different cancer types. For example, the most frequent driver genes in melanoma are *BRAF* (~50%) and *NRAS* (~30%) genes while mutation in these genes are very rare in HCC (Alkallas et al., 2020). For lung cancer, *EGFR* (47%) and *TP53* (36%) are among most frequent drivers (J. Chen et al., 2020). While *EGFR* is very rare, *TP53* is also the top driver gene in HCC (~30%) similarly in many other cancer types such as esophageal adenocarcinoma (~75%) (Deng et al., 2017), cutaneous squamous-cell carcinoma (79%%) (Y. Y. Li et al., 2015) and pancreatic adenocarcinoma (58%) (https://www.cbioportal.org/). In conclusion, mutation burden and driver landscape shows great variaability across different cancer types.

2.4.4 Novel HCC drivers are uncovered

Leveraging a large collection of HCC genomes, I have identified a number of novel drivers (Figure 2.3). Several of these novel drivers are particularly interesting. For example, *FRG1*, which was shown to increase invasiveness in breast cancer cell lines (Tiwari et al., 2019, 2017) are found to lead to extremely poor overall survival for the mutated patients (Figure 2.9). In addition, evolutionary analysis revealed that almost all mutations in this gene are subclonal, suggesting that *FRG1* might drive tumor progression

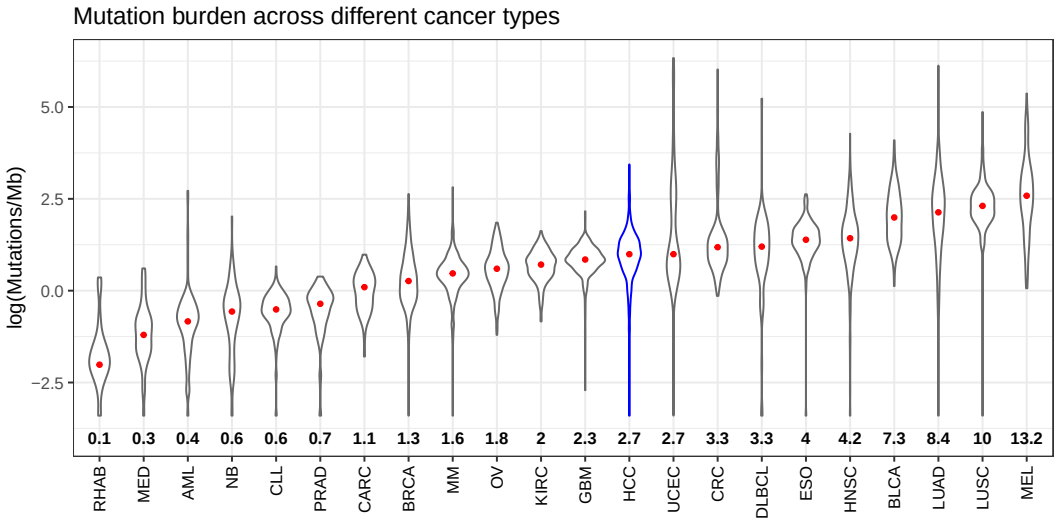


Figure 2.15: TMB across different cancer types. Number at the bottom is median number of mutations per megabase. Mutation data for other cancer types were extracted from http://www.tumorportal.org/. Mutation data for HCC is not included in the original study and mutation rate distribution of the combined HCC cohort in this study (n=1349) is shown in the plot. Cancer abbreviations: RHAB: Rhabdoid tumor, MED: Medulloblastoma, AML: Acute myeloid leukemia, NB: Neuroblastoma, CLL: Chronic lymphocytic leukemia, PRAD: Prostate adenocarcinoma, CARC: Carcinoid, BRCA: Breast, MM: Multiple myeloma, OV: Ovarian, KIRC: Kidney clear cell, GBM: Glioblastoma multiforme, UCEC: Corpus Endometrial Carcinoma, HCC: Hepatocellular Carcinoma, CRC: Colorectal, DLBCL: Diffuse large B-cell lymphoma, ESO: Esophageal adenocarcinoma, HNSC: Head and neck, BLCA: Bladder, LUAD: Lung adenocarcinoma, LUSC: Lung squamous cell carcinoma, MEL: Melanoma.

in HCC. In addition, *KCNN3* is found to be an oncogene based on the clustering of mutations in the protein. *KCNN3* encodes for SK3 protein which is a calcium-dependent potassium channel and was reported to activate proliferation in breast (Prevarskaya et al., 2010) and colorectal cancer cells (Gueguinou et al., 2017). Lastly, *DOCK2* is the other interesting novel driver. Even though, it is predominantly expressed in lymphocytes and often contributed to their migration. Given the ample functional and clinical evidences, further studies can follow up these genes for their functional and clinical evidences.

In many cancer types such as lung cancer, one of the most leading driver mutations can often been targeted (e.g. EGFR mutations in lung adenocarcinoma, BRAF mutations in melanoma). In HCC, even though a number of cancer drivers have been found, the actionable mutations are not so abundant. Sorafenib, a kinase inhibitor, is one the limited first-line treatment option for HCC. However, it is only effective for a small set of patients and only extend patient survival for 3 months (Llovet et al., 2008). This can be attributable to non-overlapping genes between known drivers of HCC and sorafenib targets such as *BRAF* and *PDGFRB*. Lack of a known targetable mutation in HCC puts the identification of novel driver genes at top priority. Identification of these novel drivers can potentially help to develop personalized drug treatments for HCC.

One special category of drivers is chromatin remodeling genes such as *ARID1A* and *ARID2* which are mutated more frequently across patients. Among the novel drivers identified in my work, *PBRM1*, *ARID1B*, *KDM6A* and *KMT2D*

are also chromatin remodelers. These new genes further confirmed the role of chromatin remodelers in HCC. One interesting finding is that drivers in chromatin remodeling pathway arises late compared to other pathways (Figure 2.13 E), similar to a number of other cancer types such as non-small cell lung cancer (Jamal-Hanjani et al., 2017). As chromatin remodeling play pivotal roles in transcriptional regulation, epigenetic reprogramming might be driving tumor progression in many HCC genomes. Understanding epigenetic regulations in tumor progression for HCC and testing epigenetic drugs for HCC management can be an interesting path for the field.

2.4.5 Clinical associations of drivers

Clinical co-occurence analysis of driver genes revealed many interesting associations. For example, *BAP1* driver mutations enriched in female patients. Interestingly, in a previous study *BAP1* mutations stratified only female patients in clear cell renal carcinoma (Ricketts & Linehan, 2015). In addition, *ALB* and *CTNNB1* are enriched in male patients. As HCC incidence is much prevalent in males, these genes might provide insights into this disparity. Another interesting link is enrichment of *TP53* mutations in HBV positive HCCs. It is known that oncoprotein of HBV virus, HBx, can bind to p53 protein and might inactivate it (Lupberger & Hildt, 2007). Lastly, *TP53* mutations are enriched in Asian patients which suggests ethnic differences in the HCC genome. One reason for higher *TP53* mutations might be aflatoxin exposure which is known to generate adduct in the binding region of p53 and HBV-*TP53* co-occurence as these etiological factors are dominant in Asia

(Razavi-Shearer et al., 2018; W. Zhang et al., 2017).

2.4.6 Timing of drivers can help design better treatments

Timing driver genes is crucial to understand how tumors evolve and to design better treatment strategies. McGranahan et al. (2015) identified many subclonal driver mutations including one of the targetable gene, *IDH1*, in glioblastomas highlighting the importance of timing of driver genes when designing treatment schemes. Even though HCC currently does not have a known targetable mutation, understanding the evolutionary landscape will help to identify drug targets. For example, drug screening to target early driver mutations might help to obtain better results. In this study, timing analysis of HCC drivers revealed the important evolutionary landscape of HCC.

Even though a large sample size was employed in the integrative study (n=1349), there is still a large proportion of patients without any identified driver gene (~15%). Increased statistical power provided by the large sample size allowed us to identify more drivers compared to the previous studies. However, saturation analysis revealed that the available sample size is not enough to identify rare drivers (Figure 2.7)(Lawrence et al., 2014). Since my current analysis only focused on the protein coding regions in the genome as majority of available HCC samples are sequenced with WES technology, many other driver events such as non-coding changes in regions such as the 5' UTR region of TP53 gene and 3' UTR region of the ALB gene (Juul et al., 2017; Rheinbay et al., 2020) can contribute to the evolution of HCC. Combining driver events across multiple layers can provide a more holistic landscape of

driver events in HCC.

3 Chapter 3: Ethnic comparison of hepatocellular carcinoma

3.1 Introduction

Across the world, the incidence of HCC is rather heterogeneous. Based on 2018 GLOBCAN statistics, around 70% of all cases across the world were reported in Asia. Furthermore, many cases in Asia (75%) are condensed in the East Asia (Bray et al., 2018). Moreover, the disease etiology and clinical outcome of HCC are also quite different geographically (Rich et al., 2019; Villanueva, 2019). Despite rapid progress in understanding HCC genomes individually in each cohort, there are still significant gap in the field.

Firstly, given diverse etiological backgrounds in HCC, ethnic comparison characterizing divergent molecular events specific to each etiological background has only be explored individually in different aspects, without a systematic comparison, including mutational signatures (e.g. dominant T>A changes in Asians) and transcriptomic subtypes (e.g iCluster1 enriched in Asians) (Ally et al., 2017; Chaisaingmongkol et al., 2017; Totoki et al., 2014). Secondly, in addition to inter-patient differences, cancer also show extensive heterogeneity within a tumor. Using pan cancer datasets, several recent studies reported varying levels of intra-tumor heterogeneity (ITH) and have found significant correlation between ITH and clinical outcome across tumor types (Andor et al., 2016). However, HCC was not included in the large-scale pan-cancer ITH studies possibly due to unavailability of the HCC data at the time. In HCC,

recent studies reported different degrees of ITH, but these analyses were limited to small cohorts with limited number of patients (Friemel et al., 2015; Zhai et al., 2017). Thus, clinical consequence of HCC as well as racial disparities in terms of ITH is understudied.

One of the best datasets for ethnic comparison is the TCGA-LIHC cohort. It comprised equal amount of Asian (n=161) and European patients (n=187). More importantly, the data collection and sequencing protocols are uniformly conducted. In this chapter, I will first carry out a systematic comparison across multiple layers including; 1) Patient clinical phenotypes, 2) Tumor mutation burden (TMB) and driver gene frequencies, 3) Mutational signatures, 4) Copy number variation (CNV), 5) Intra-tumor heterogeneity, 6) integrative survival models. I aim to provide a holistic view of ethnic differences in HCC. Systematic comparison revealed higher genomic instability in Asians with a series of molecular events segregating differently between the two ethnic backgrounds. Most strikingly, I identified a clinically aggressive RNA subgroup unique to Asians, driven by multiple ethnic specific genomic alterations. Through evolutionary analysis, I found multiple ITH features can provide important prognostic value to patient survival and predictive survival models perform much better in Asians possibly driven by ethnic specific molecular events. For the first time, I conducted a comprehensive integrative analysis of HCC genomes and provide novel insights into ethnic differences in HCC.

3.2 Materials and Methods

3.2.1 Driver differences in Asians and Europeans in the TCGA cohort

WES data from the TCGA cohort was downloaded from GDC and somatic mutations were called using mutect. Among the identified driver genes (n=62), Fisher's Exact test was applied to test for frequency differences for each gene (limiting to genes with at least three occurance) and Benjamini-Hochberg method was used for multiple hypothesis testing. Drivers with q value less than 0.1 were selected as significantly different genes.

3.2.2 Identification of signature groups

A list of 10 liver related mutational signatures from previous studies were collected (Letouzé et al., 2017; Schulze et al., 2015). DeconstructSigs was used to decompose the mutations into these signature groups (Rosenthal et al., 2016). COSMIC v3.1 signatures was used for the deconvolution (PCAWG Mutational Signatures Working Group et al., 2020). After deconstruction, signatures with mean proportion greater than 2% or a maximum proportion of 20% were kept and deconstruction step was repeated with the remaining signatures. With the contribution of different signatures estimated for each patient, signature proportions were clustered using the hierarchical clustering algorithm with the Euclidean distance and 'ward.D' method in R. Timing of signatures was done by decomposing early and late mutations based on the CCF threshold using deconstructSigs separately. Comparison of signatures between early and late

tumorigenesis was done using the paired Wilcoxon test.

3.2.3 Copy number analysis

Sequenza was employed to infer the integer copy number using the WES data downloaded from the GDC (Favero et al., 2015). Genomic instability index (GII) was calculated by comparing the copy number of each segment with median copy number across the genome of the patient. GII is simply the fraction of the genome with an integer copy number different from the median ploidy. Somatic copy number alteration (SCNA) score was calculated for arm and broad scale CNV output from the GISTIC algorithm by adopting method from Yuan et al. (2018). Based on the amplitude of events, if the CNV value is greater than 1 or less than -1, 2 and -2 weights were assigned respectively. For values between (0.25, 1) and (-0.25, -1), 1 and -1 were used as the weights. CNV values between 0.25 and -0.25 were weighted as 0. Then, for each patient, absolute values of weighted values are summed up for focal (SCNA focal) or arm level (SCNA arm) values separately. Finally, scores were rank-normalized in order to eliminate the effect of outliers. GISTIC (Mermel et al., 2011) was employed to identify significantly purturbed CNVs for the Asian and European patients from the TCGA cohort separately (with "-genegistic 1 -smallmem 1 -broad 1 -brlen 0.5 -conf 0.95 -armpeel 1 -savegene 1 -gcm extreme"). Arm level frequencies were compared across cohorts using the results from the broad_significance_results.txt based on the Fisher's exact test. Multivariate logistic regression was conducted using the glm() function in R (with "binomial" family) and p-value for each variable was adjusted for

multiple testing using the Benjamini-Hochberg method and 0.1 was used as the cut-off for significance. For focal amplifications, q values from the scores.gistic output was used. Common and private peaks were identified by overlaping peak limits from the Asian and European cohorts using the "GenomicRanges" R package. Driver genes, a list of pan-cancer amplifications and deletions49 and a list of known liver copy number events27 were labeled for the peaks. Low purity samples were first filtered away before GISTIC analysis (purity<0.25, n=10) for quality purposes.

3.2.4 RNA-Seq analysis

Raw gene counts were downloaded from GDC (https://portal.gdc.cancer.gov/). Protein coding genes was used for further analysis and lowly expressed genes were filtered out (i.e. removing genes with less than 5 counts in at least ten patients). Expression levels were normalized by estimating size factors from DESeq2 and normalized counts were subsequently log2 transformed after adding 1 pseudo count.

For molecular subtypes, top 3000 most variable genes (based on median absolute deviation, MAD) were selected for both Asian and European cohorts separately. Non-negative matrix factorization (NMF) algorithm was applied by running NMF R package with the Brunet algorithm50. Number of ranks from 2 to 6 were iteratively run for 200 times. Optimal rank (number of subtypes) was selected using the highest cophenetic correlation and the highest consensus silhouette values.

Mapping of homologous subtypes was conducted using SubMap method from

the Genepattern with default parameters31. Differentially expressed genes were identified using DESeq2 (Love et al., 2014). Gene set enrichment analysis was conducted using the "fgsea" method51. Hallmark (v6) and C2_CGP (v7) gene sets were used for fgsea and genes were ranked according to a combined score (sign of log fold change times -log10 of p-value). Significant pathways were extracted if the fdr is less than 0.05.

In order to measure the enrichment of a pathway in a group of patients, gene set variation analysis (GSVA) was conducted to calculate a pathway level score 54 . Comparison of clinical and genomic feature differences were conducted using the Wilcoxon test for continuous variables and Fisher's Exact test was used for categorical variables. P-values were adjusted using the Benjamini-Hochberg (fdr) method (cutoff of 0.1).

3.2.5 Calculation of ITH metrics

Percentage of late mutations (pLM) was calculated by dividing the number of late mutations (CCF <0.8) by the total number of mutations in each tumor. MATH score was calculated as described in the original study (Mroz et al., 2013). Pyclone was employed to infer the clonal structure of the tumor (Roth et al., 2014). Binomial density and 10,000 iterations were selected for the MCMC (the first 1000 iterations were treated as the burning phase). Shannon index was calculated using the number of mutations in each subclone identified by pyclone as:

$$SI = -\sum_{i=1}^{s} p_i \log p_i$$

Where s is the number of subclones and p is the mean CCF of each cluster.

3.2.6 Selection of features for integrative survival analysis

A total of 39 features from clinical (n=5), molecular (n=16), driver (n=6) and ITH (n=6) categories were compiled. Clinical category included classical features including gender, age, stage, HBV and HCV status. Molecular category included: 1) generic generic tumor features including purity, ploidy, RNA subtypes, TMB, SCNA; 2) proportions for common mutational signatures (i.e. with a mean proportion of no less than 5% across patients which include SBS4, SBS5, SBS12, SBS22); 3) Immune features including immune subtypes, TIDE score, MDSC score, GEP, TAM_M2 score, CAF score and total tumor infiltrating lymphocytes (TILs) score. Driver category included driver genes with at least 15 mutations across all patients. Finally, ITH category included pLM, MATH score and Shannon's index as well as their categorized version as low, medium and high. This categorical version was included to account for possible non-linear relationship between survival and the dependent variable.

Univariate Cox survival analysis was applied to each feature and features which stratify patients ($p<0.05$) in either Asian, European or the combined TCGA cohort were used for further analysis. The best subset of predictive features was selected using stepwise Cox regression with both forward and backward selection

(bi-directional). For ranking of variables based on relative importance, the final set of variables were fit in a multivariate Cox model and likelihood ratio test (LRT) was applied using Anova() function in R. 80% of Asian and European cohorts were randomly sampled 50 times and importance was calculated based of proportion of chi-squared values from likelihood-ratio test in each resampled dataset The mean value of importance was shown in Figure 3.35. Comparison of survival prediction accuracy was conducted by splitting individual cohorts as 80% training and 20% test set followed by fitting either the all features in the final subset or features in individual categories (clinical, molecular, Driver, ITH) using training set (80%) and calculating concordance index (c-index) using the test set (20%).

3.3 Results

3.3.1 Patient cohort characteristics

In this Chapter, I used TCGA cohort for ethnic comparison of Asian and European HCC. The reasons for using TCGA for ethnic comparison are: 1) Accesibility to the raw sequencing data for multi-layer comparison (e.g. Copy number comparison, RNA subtypes), 2) balanced number of Asian and European patients and 3) uniform sequencing technologies suggesting minimum batch effect. TCGA ethnic comparison cohort consists of 348 patients including 161 Asian and 187 Europeans. While 67% of patients are male, only 33% of patients are female. There are 190 non-viral cases (NBNC), 96 HBV+ cases, 38 HCV+ cases. Furthermore, 6 patients have multi-viral

HBV and HCV infection. Majority of tumors have TNM stage I and II (49%, 24%) and only 1% of tumors is stage IV. Finally, the median age of patients in the cohort is 61.

3.3.2 Ethnic differences in clinical phenotypes and mutational landscape

In order to systematically survey the difference between the Asian and European patients, I first compared the clinical characteristics of patients in Asian and European patients (Denoted as TCGA-Asian and European). The most significant difference is the viral status of patients (p = 6.42e-31, Figure 3.1 A). While around 60% of Asian patients are HBV positive, only 4% of Europeans are HBV carriers. Most of the European patients are non-viral cases (76%) with viral cases being mostly HCV carriers (18%). On the other hand, HCV cases is rather rare (4%) in the Asian cohort. In addition to viral status, European patients have a relatively higher proportion of female patients (44% vs 21%, p = 7.79e-06, Figure 3.1 B) and older age at diagnosis (median age 66 vs 55, p = 3.66e-12, Figure 3.1 C). In general, the two cohorts are quite similar in other clinical phenotype including tumor stage, microvascular invasion status (MVI), alpha-fetoprotein (AFP) levels as well as tumor purity (Figure 3.2).

Through the clinical comparison, a relatively similar profile is observed. I was curious whether the two cohorts will differ in the genomic landscape. By processing the raw sequencing data downloaded from GDC and subsequently process them using in-house pipeline (Methods), I found that Asian

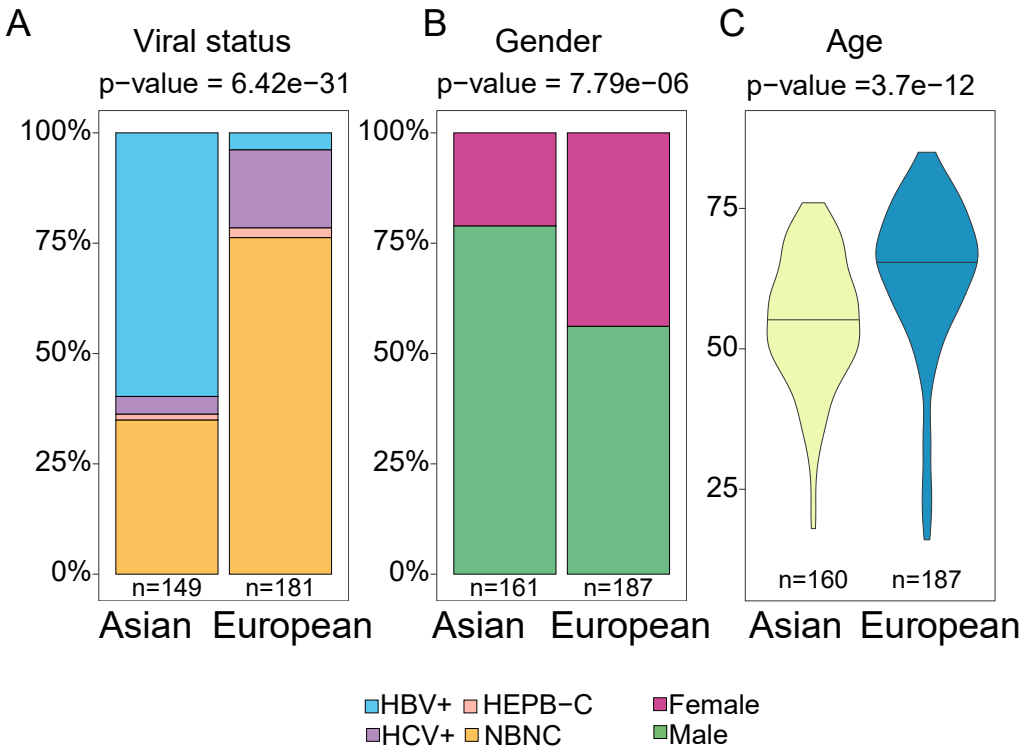


Figure 3.1: Significantly different clinical phenotypes. (A) Comparison of viral status. NBNC stands for Non B/Non C (non viral) and HEPB-C is short for Hepatitis B and Hepatitis C multiviral infection. While majority of Europeans are non-viral cases (NBNC), Asian patients are mostly HBV positive. (B) Comparison of gender proportions. Female patients are in higher proportion in the european cohort. (C) Comparison of patient age. The median patient age is significantly higher in Europeans.

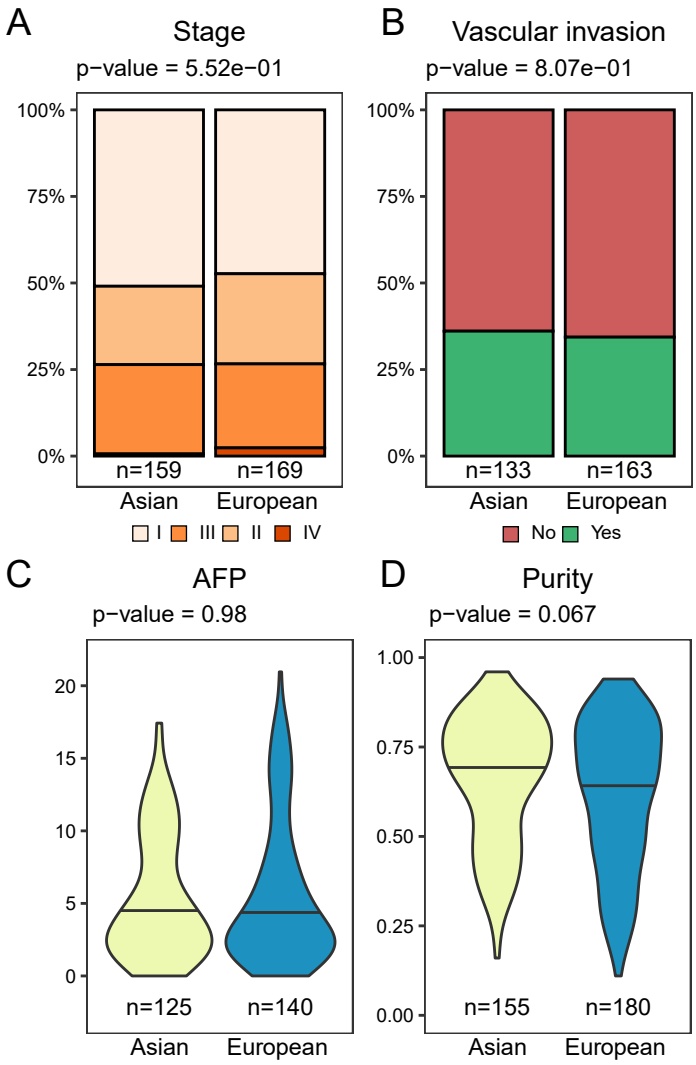


Figure 3.2: Similar clinical feature and purity plots. (A) Tumor stages are in similar proportions between two cohorts. (B) Both cohorts have similar proportion of tumors with microvascular invasion which is a predictor of tumor invasiveness in HCC where tumor cells invade blood vessels. (C) Alpha-feto protein levels are similar between cohorts. (D) Tumor purities (fraction of tumor cells in the bulk sample) are also similar between two cohorts.

patients have significantly higher TMB (4 mutations/megabase vs 3.5 mutations/megabase, p = 9.90e-03 Figure 3.3 A). Since many variables including clinical phenotypes might correlate with mutational load of the tumor (e.g. age), a multivariate linear model taking into accout many of these factors was constructed. TMB difference between two ethnic groups remains significant after controlling for clinical variables and tumor purity (p = 4.58e-03, Figure 3.3 B). Thus, significantly higher TMB was observed in Asians.

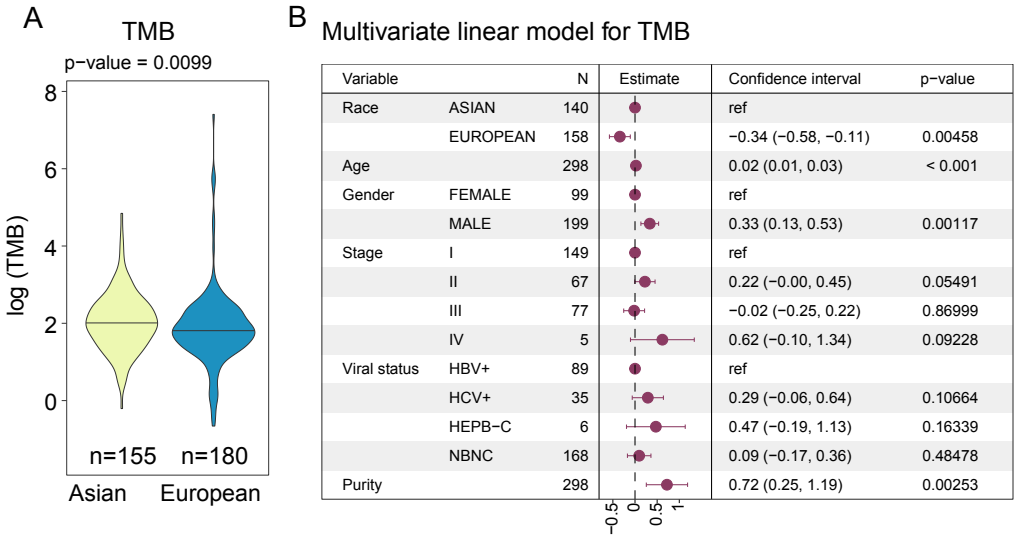


Figure 3.3: Comparison of the tumor mutation burden (TMB) between Asian and European cohorts (A) A higher TMB is observed in the Asian cohort. Wilcoxon's test wass applied for p-value calculation and log2 transformation was implemented for plotting. (B) The race variable is still siginicant after adjusting for clinical features and tumor purity.

Given the higher TMB observed in the Asian cohort, I asked whether driver genes will also segregate differently in two ethnic backgrounds. When I compare frequencies of driver genes (n=62, Chapter 2) in Asians and Europeans, *TP53*,

RB1, *EEF1A1* and *CDKN2A* had significant differences between the two cohorts at the nominal cutoffs (p-value <0.05, Table 6). After multiple test correction, only *TP53* and *CDKN2A* remained significant (q-value = 0.002 and 0.07 respectively, Figure 3.4). Among all five driver genes, most of the driver genes have higher frequencies in the Asian cohort with the exception of *EEF1A1*.

Table 6: Driver freqeuency comparison between Asian and European cohorts and p-values

Gene	Asian_percentage	European_percentage	p_value	q_value
CDKN2A	6.451613	0.000000	0.0003828	0.0164623
TP53	36.774194	23.497268	0.0037896	0.0814762
EEF1A1	1.290323	7.650273	0.0245425	0.2638318
RB1	10.256410	4.395604	0.0225475	0.2638318

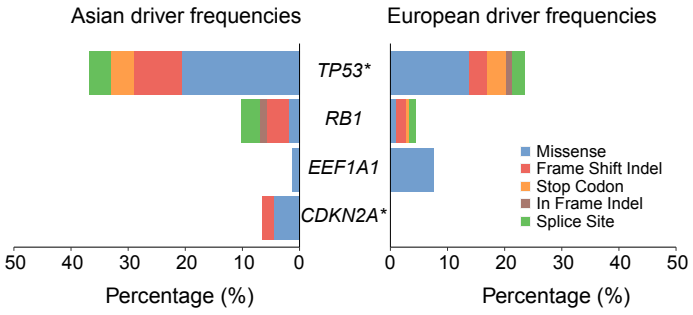


Figure 3.4: Driver genes with significantly different frequencies between Asian and European cohorts. Different color represents the percentage of mutations for each mutation type. Fisher's Exact was was used to test for frequency difference between two cohorts. The star indicates a q-value of less than 0.1. Only TP53 and CDKN2A remains sigfnificant after multiple testing correction.

The mutational analysis revealed an overall increase in TMB specific to Asians with a small subset of drivers more frequent in Asia. I then asked

whether certain mutational process specific to Asians can explain the observed differences. For example, the intake of Traditional Chinese Medicine (TCM) herbs containing aristocratic acid (AA) was found in multiple cancer types including liver cancer (Ng et al., 2017). Several other studies have identified a list of signatures known to drive HCC tumorigenesis (n=10) (Letouzé et al., 2017; Schulze et al., 2015). These signatures included clock-like age-related signatures (SBS1 and SBS5), smoking (SBS4), DNA mismatch repair (SBS6), aristolochic acid (AA) and aflatoxin B1 exposure (SBS22 and SBS24), liver cancer associated signatures (SBS12 and SBS16) and two other signatures with unknown etiology (SBS17 and SBS23). To identify contributions of these signatures across the cohorts, I used deconstructSigs to project the mutations into the contributions of these signatures (Rosenthal et al., 2016). 9 out of 10 signatures (SBS1 (4%), SBS4 (12%), SBS5 (55%), SBS6 (2.4%), SBS12 (6%), SBS16 (3%), SBS22 (5%), SBS23 (1.6%), SBS24 (3.6%) have appreciable proportions in the combined Asian and European cohort (mean proportion > 2% or max proportion >20%, Table S8).

In order to compare the ethnic differences in individual signatures, I compared signature proportions between Asian and European patients (Figure 3.5). The most significant differences are observed for aging signature SBS5 (p= 4.7e-08) and AA related signature SBS22 (p= 1.2e-08). While aging signature SBS5 is higher in Europeans, AA related signature SBS22 signature was enriched in Asians. While higher SBS22 signature in Asian could be because of widespread TCM herb usage (Ng et al., 2017), high amounts of aging signature could be attributable to patients with older age in European cohort as shown before in

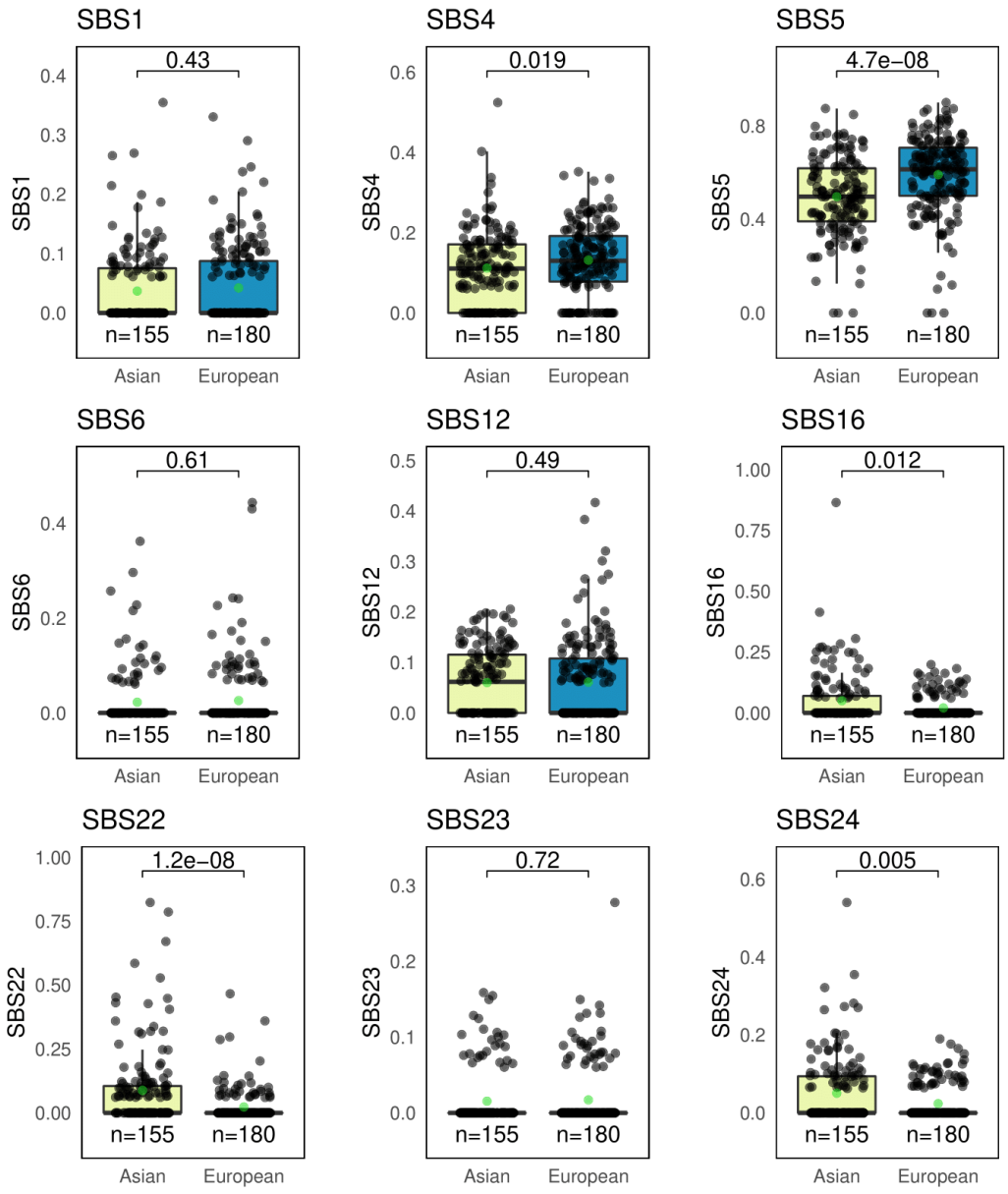


Figure 3.5: Comparison of proportions for 9 signatures between Asian and European cohorts. Signature proportions obtained from deconstrucsigs were compared with Wilcoxon's rank sum test. p-values are presented on top of boxplots.

Figure 3.1. In addition to SBS5 and SBS22, smoking signature SBS4 (p= 0.019, higher in Europeans) and aflatoxin exposure related signature SBS24 (p= 0.005, higher in Asians) show marginal differences between two cohorts (Figure 3.5).

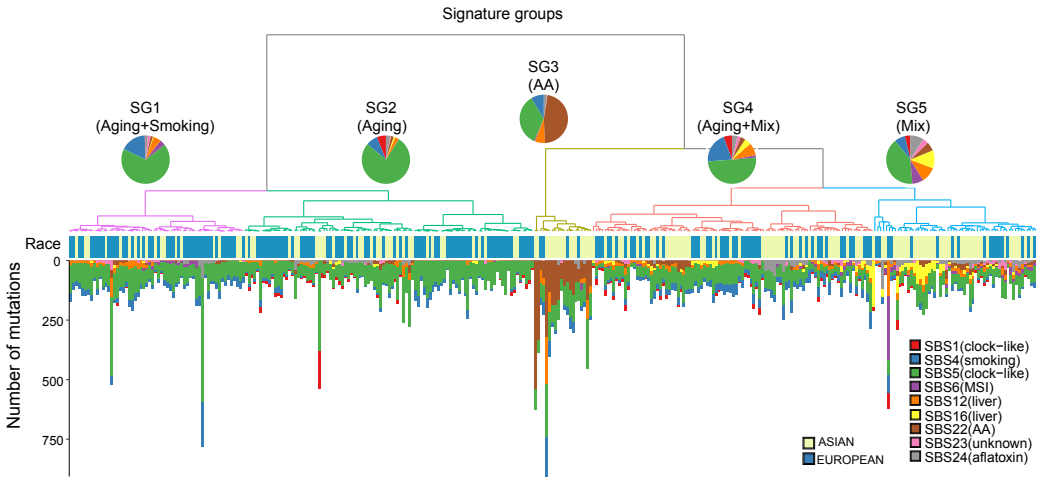


Figure 3.6: Signature groups across patients were obtained using hierarchical clustering of signature proportions. SG3 is a distinct signature group which is dominated by aristolochic acid (AA) signature SBS22 and mostly consist of Asian (race annotation in the middle) and high TMB patients (bottom barplot).

To further interrogate the ethnic differences, patients were clustered into five sub-groups using the proportion of different signatures (Figure 3.6). Groups SG1 and SG2 are dominated by SBS5 (clock like) signature and are enriched for European patients (p = 1.15e-06, Figure 3.7 A). SG3 is the group of patients with strong the AA signature and much higher TMB than other groups (Figure 3.7 B) . SG4 has a dominant signature SBS5 together with an appreciable proportion of SBS4 (smoking) and a mix of other signatures. SG5 has much higher frequency of liver related signatures (SBS12 and SBS16) and are more enriched for Asian patients (p-value= 1.15e-06, Figure 3.7 A).

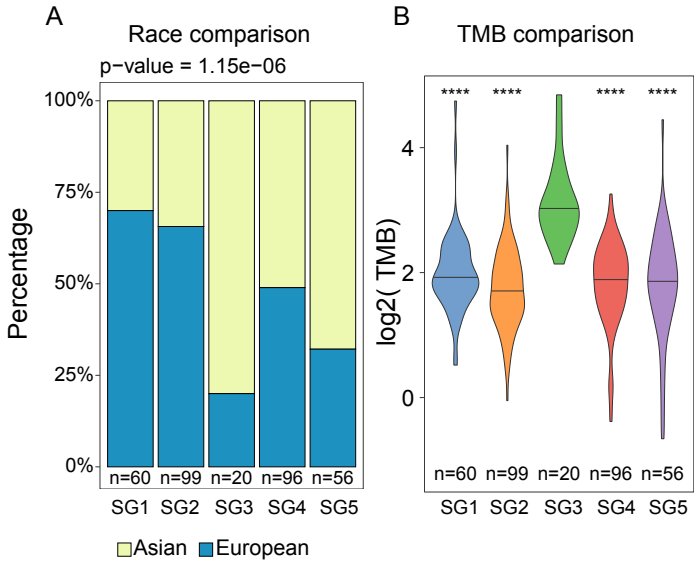


Figure 3.7: Correlations of signature groups. (A) Proportions of Asian and European patients across signature groups. (B) Comparison of TMB across signature groups. SG3 has the highest tumor mutation burden compared to all other signature groups. SG3 reference group in the Wilcoxon test and p-value <= 0.001 were labeled as "****".

Multivariate linear model for TMB with AA and *TP53* mutation status

Variable		N	Estimate	Confidence interval	p−value
Race	ASIAN	140		ref	
	EUROPEAN	158		−0.06 (−0.29, 0.16)	0.57322
Age		298		0.02 (0.01, 0.02)	< 0.001
Gender	FEMALE	99		ref	
	MALE	199		0.31 (0.13, 0.50)	< 0.001
Stage	I	149		ref	
	II	67		0.17 (−0.04, 0.38)	0.11341
	III	77		−0.03 (−0.25, 0.18)	0.76426
	IV	5		0.65 (0.00, 1.30)	0.04979
Viral status	HBV+	89		ref	
	HCV+	35		0.15 (−0.17, 0.47)	0.36305
	HEPB−C	6		0.38 (−0.22, 0.98)	0.21638
	NBNC	168		0.03 (−0.21, 0.27)	0.79456
Purity		298		0.56 (0.14, 0.98)	0.00983
SBS22		298		2.70 (1.98, 3.42)	< 0.001
TP53	wt	211		ref	
	mut	87		0.33 (0.14, 0.51)	< 0.001

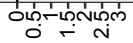


Figure 3.8: Multivariate linear regression for TMB and covariates including AA and TP53. In this model, TP53 and AA signature (SBS22) was also added to the model in addition to clinical variables and tumor purity. TMB difference between Asian and European cohorts is not significant anymore.

As previously noted that Asians have higher TMB compared to Europeans (Figure 3.3). In addition, higher frequency of *TP53* driver gene frequency was observed in Asians (Figure 3.4). Higher proportions of Asian patients in SG3 which has much higher number of mutations seem to correlate with the higher TMB in Asians. To explore variables that might explain the higher TMB in Asians, I generated an integrated multivariate linear model by adding AA signature proportion and TP53 mutation status to the model (Figure 3.8). Interestingly, both AA signature and the TP53 mutation status are significantly associated with TMB ($p<0.001$) and ethnic difference in TMB was no longer significant after controlling for these variables. Interestingly models with AA signature and TP53 mutation status fit the data significantly better than the baseline model without these variables (i.e. likelihood ratio test, residual sum of squares (RSS) 172 vs 140, p = 4.63e-15). This implies that higher TMB in Asians is mainly due to higher proportions of AA signature and higher frequency of TP53 mutations in Asians.

To understand the timing of these mutational events along the history of tumorigenesis, fractions of cells carrying that mutations can be used as an estimate, reasoning that early mutations will be carried by all cancer cells. I calculated the cancer cell fractions (CCF) for all mutations and decomposed signatures separately for early (CCF>=0.8) and late mutations (CCF<0.8). Interestingly, the proportion of smoking and AA signatures were significantly lower in the late stage of tumorigenesis compared to early mutations (Figure 3.9). On the other hand, MSI and liver associated signatures SBS6 and SBS12 have higher proportions in the late stages of tumorigenesis

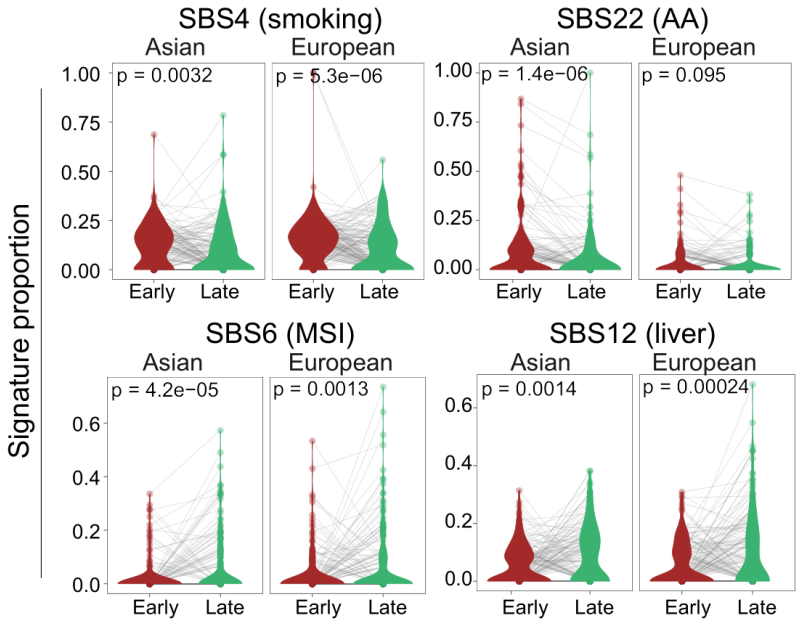


Figure 3.9: Signatures with significant change in proportions between early and late mutations. After mutations were categorized as early and late based on the cancer cell fractions (CCF), signature proportions are calculated separately and compared using paired Wilcoxon's rank sum test. While SBS4 and SBS22 proportions are lower at the later stage of tumorigenesis (i.e late mutations) compared to early stages, an increase is observed for SBS6 and SBS12. These patterns are similarly significant between both cohorts except that SBS22 difference is not significant in the European cohort possibly due to much lower proportions of this signature in Europeans.

(Figure 3.9). These patterns are common across Asian and European cohorts with the exception of AA signature which possibly did not reach significance due to much lower proportions of this signature in European cohort ($p = 0.095$, Figure 3.9). It is important to notice that other signature groups did not show any difference along the history of tumorigenesis (Figure 3.10).

3.3.3 Ethnic comparison of copy number landscape

In addition to point mutations, driver genes and mutational signatures, copy number alterations (CNAs) are the other important mutational events driving tumorigenesis. Since the effect of copy number alterations depends on both the frequency and the magnitude of copy number alterations, I calculated a somatic copy number alteration score (SCNA) integrating these factors for broad and focal level CNAs separately (See Methods). In addition, CNV burden can be represented as genomic instability index (GII) which is simply the altered genome fraction regardless of the magnitude of alteration. Interestingly, similar to TMB, arm level SCNA scores were also significantly higher in Asians ($p = 3.60e\text{-}04$, Figure 3.11 A). Moreover, higher genome instability stays true even when we ignore the magnitude of CNV changes and only compare GII across the two cohorts (Figure 3.11 B).

As discussed earlier (Figure 3.8), TP53 point mutations is significantly associated with higher TMB and differences between Asian and Europeans in TMB were not significant after controlling for other covariates. A significant association between SCNA and TP53 mutation was also observed in both Asian and European cohorts (Figure 3.12 A). Given the higher frequency

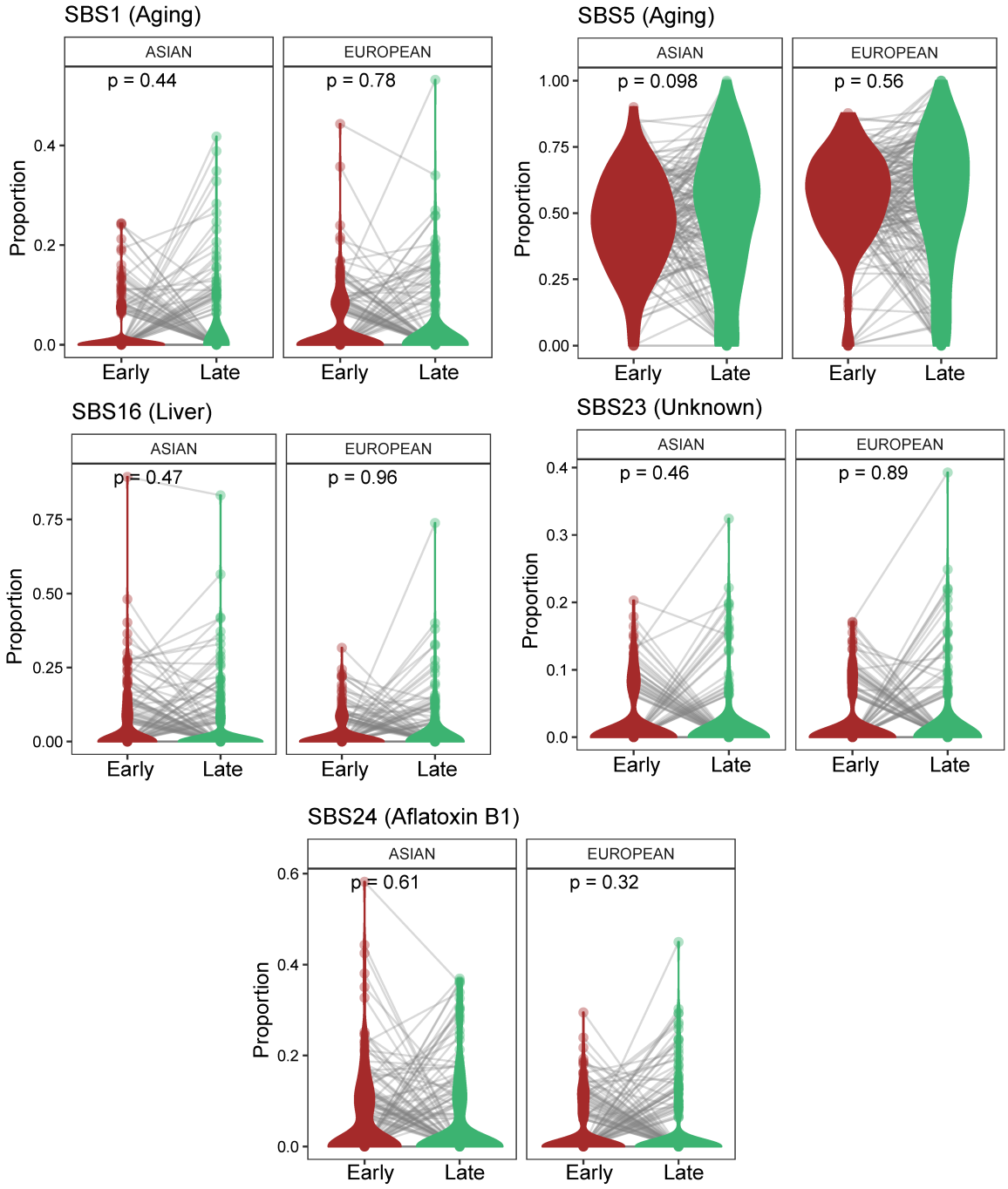


Figure 3.10: Signatures with similar proportions at early and late mutations. After mutations were categorized as early and late based on the cancer cell fractions (CCF), signature proportions are calculated seperately and compared using paired Wilcoxon's rank sum test.

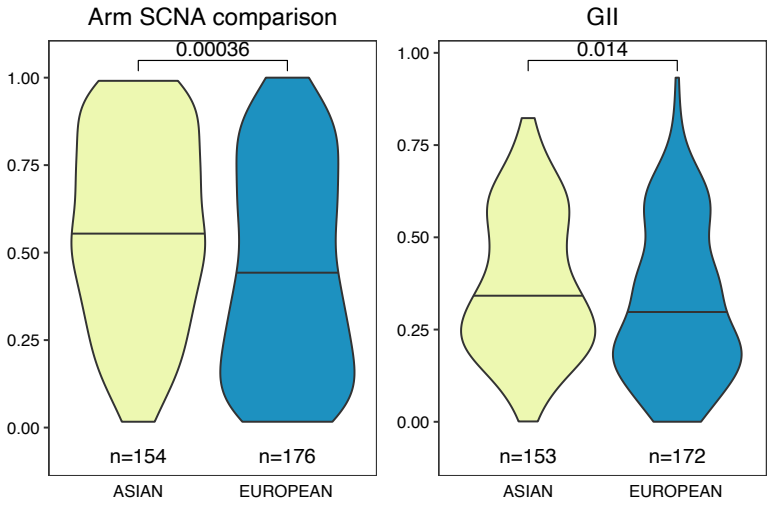
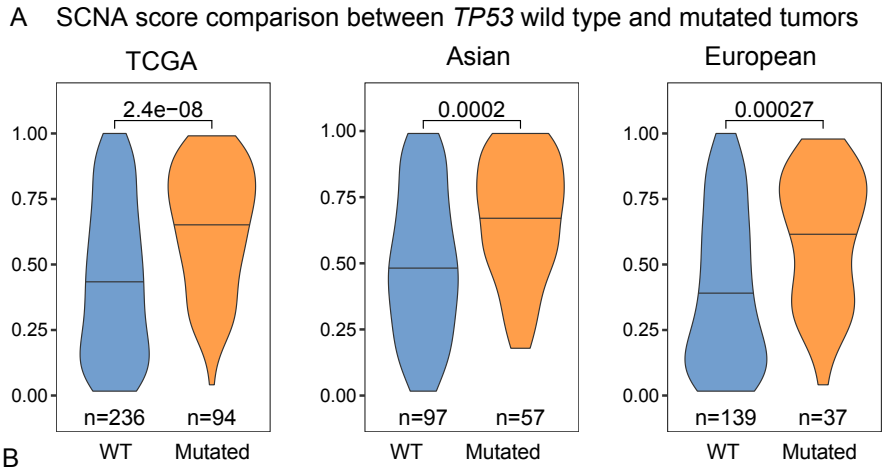


Figure 3.11: Comparison of copy number variation (CNV) burden using (A) somatic copy number alteration (SCNA) score and (B) Genomic instability index (GII). Horizontal lines on the violin plots mark the median value of the distribution. SCNA score is similar to GII but it also takes the magnitude of CNV event into account in addition to altered genome fraction. Asian genomes are more instable using both metrics.

of TP53 mutations in Asians, a multivariate linear model including clinical covariates, tumor purity as well as TP53 mutation status was generated. After controlling for covariates, the ethnic difference in arm level SCNA scores remained significant ($p = 0.02$, (Figure 3.12 B)).

In order to understand the contribution of individual chromosomes to the overall SCNA scores, I compared the ethnic differences in arm level CNV events for each chromosome arm (Figure 3.13). While most of the chromosome arms have similar frequencies, 10 arms (4 amplifications and 6 deletions) were altered at significantly different frequencies (Fisher's Exact test q-value <0.1). These differential alterations included events such as 8q amplification and 4q deletions which have been reported as frequent CNV events before (Shibata & Aburatani, 2014).

Despite the arm level differences, when we compare focal CNAs using focal SCNA score, there is no statistically significant difference between two cohorts ($p = 0.32$, Figure 3.14). When plotting the focal event peaks using the GISTIC algorithm, I identified several common amplification and deletion peaks between Asian and Europeans such as *TERT* and *FGF19* amplification and *AXIN1* deletions. In addition to common peaks, Europeans do have a few private peaks including an amplification peak with *MET* gene and a deletion peak at the *BAP1* gene (Figure 3.15, Table S9). Thus, focal copy number alterations are relatively similar.



Variable		N	Estimate	Confidence interval	p−value
Race	ASIAN	140		ref	
	EUROPEAN	155		−0.10 (−0.18, −0.02)	0.0203
Age		295		0.00 (−0.00, 0.00)	0.1510
Gender	FEMALE	99		ref	
	MALE	196		−0.03 (−0.10, 0.04)	0.3577
Stage	I	148		ref	
	II	65		0.08 (−0.00, 0.16)	0.0565
	III	77		0.07 (−0.01, 0.15)	0.0866
	IV	5		0.22 (−0.03, 0.46)	0.0874
Viral status	HBV+	89		ref	
	HCV+	35		0.00 (−0.12, 0.12)	0.9975
	HEPB−C	6		−0.03 (−0.26, 0.19)	0.7632
	NBNC	165		0.01 (−0.08, 0.10)	0.8687
Purity		295		0.28 (0.12, 0.45)	<0.001
TP53	wt	211		ref	
	mut	84		0.17 (0.10, 0.24)	<0.001

Figure 3.12: Association of SCNA levels with TP53 mutations. Horizontal lines on the violin plots mark the median value of the distribution. (A) Comparison of SCNA levels between TP53 wild type (WT) and mutated patients in both Asian and European cohorts. (B) Multivariate linear model for predicting arm level SCNA levels. While TP53 mutation is significantly liked to SCNA levels, Asians still have significantly higher SCNA levels after adjusting for clinical features, tumor purity as well as TP53 mutations.

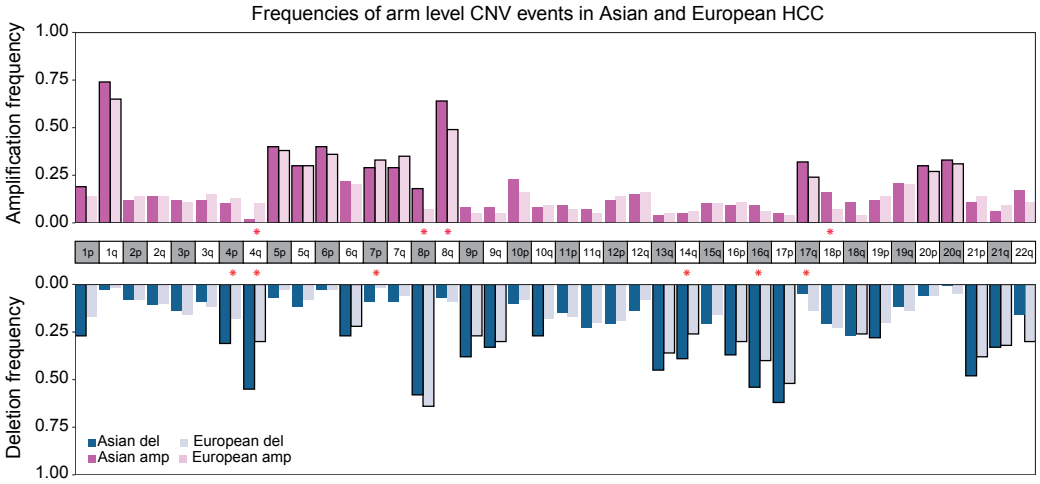


Figure 3.13: Comparison of arm level event frequencies across all chromosome arms. Frequency of deletions (Asian:navy, European:light blue) and amplifications (Asian:magenta, European:pink) of each chromosome arm are shown as bars. Fisher's exact test was applied for each arm to compare proportions between cohorts. Arms with a significant p-value was indicated with a red star near to chromosome arm labels. Black borders around the bars indicate that the bordered arm is a significantly altered arm within the cohort based on GISTIC algorithm.

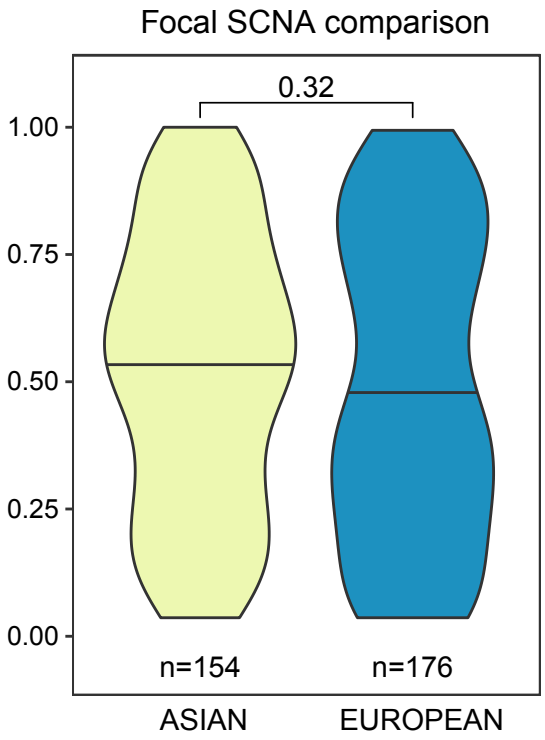


Figure 3.14: Comparison of focal SCNA levels between Asian and European cohort. No significant difference is observed in focal SCNA levels between two cohorts.

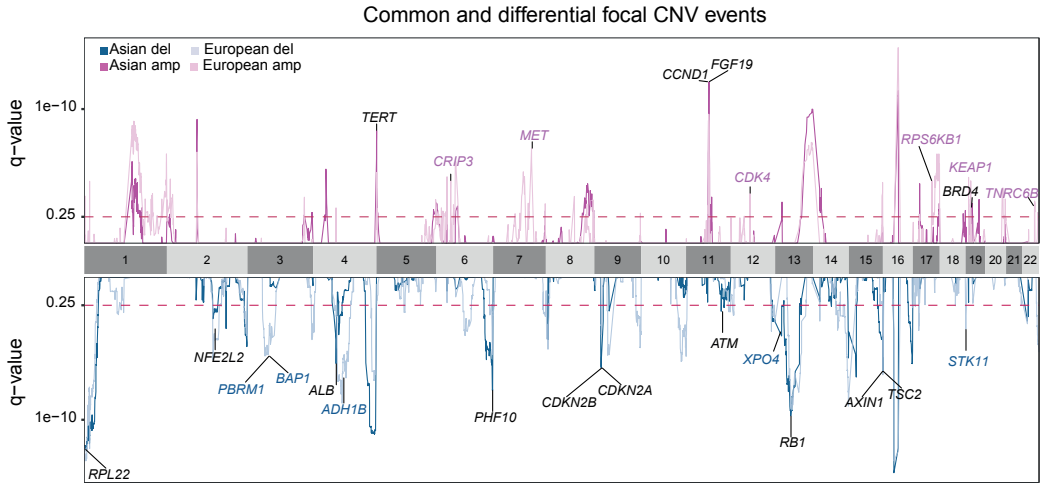


Figure 3.15: Comparison of focal CNV peaks between Asian and European cohorts. Driver genes and recurrently altered gene list was labeled if they are overlap with peaks. Genes in common peaks are labeled as black and genes at private peaks were labeled with the color of each individual cohort. While majority of the peaks are common, European cohort has a few provate peaks.

3.3.4 Ethnic comparison of transcriptomic subtypes

In addition to differences at the genomic level, transcriptomic differences between Asian and Europeans are often explored separately in each cohort without a thorough comparison across cohorts. For example, Hoshida et al 2009 described one well-recognized subtyping of HCC including two subtypes with bad prognosis (S1 and S2) and one with good prognosis (S3) Hoshida et al. (2009). However, they included a mixture of Asian and European patients for the discovery of molecular subtypes. Boyault et al. identified six molecular subtypes (G1-G6) using patients of European descendent Boyault et al. (2007) with varying levels of overlap with other RNA subtype. Even though, ethnic differences were explored in a recent study based on a Thai

cohort, the main focus was identifying a common subtype between HCC and ICC Chaisaingmongkol et al. (2017). Thus, ethnic comparisons in molecular subtypes are significantly understudied for HCC. Thus, a comparison of transcriptomic subtypes using datasets generated with the same sequencing protocol (i.e. TCGA Asian and European cohorts) can holistically survey ethnic differences in transcriptomic subtypes.

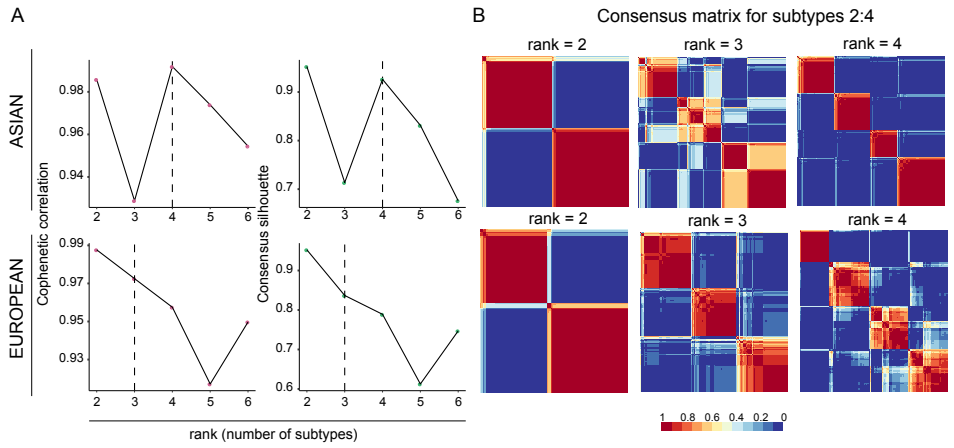


Figure 3.16: Clustering stability metrics across ranks (number of subtypes). (A) Cophenetic correlation and silhouette values across the ranks 2:4. The optimum number of subtypes are 4 and 3 for Asian and European cohorts respectively (two is not considered). (B) Consensus matrix that that shows robustness of subtypes. A coefficient of 1 (red color) implies that patients were assigned to the same subtype across multiple runs of NMF (n=200). Top panels are results for Asian cohort and the bottom panels are for European cohort.

Using non-negative matrix factorization (NMF), I identified two or four optimal number of subtypes for Asian cohort and two or three subtypes for European cohort using both cophenetic correlation coefficient and sillhouette values (Figure 3.16 A) as well as consensus matrix (Figure 3.16 B). When I

split both cohorts into two subtypes, the transcriptomic landscape matched significantly using subtype mapping method SubMap (Hoshida et al., 2007) (Figure 3.19). Further comparison of pathways between the two subtypes revealed one subgroup with up-regulated cell cycle pathways (e.g. "G2M checkpoint"), but down regulation of metabolic pathways typical to liver function (e.g. "Bile acid metabolism") in both cohorts (Figure 3.20 A, Figure 3.21 A, B). This matches a general trend across cancer types that there is a relatively "benign" group with transcriptomic profiles similar to the normal tissue, but a "malignant" group with upregulated cell cycle pathways. Despite the functional similarity in the basal split across cohorts, the two-subgrouping only stratifies patients for overall survival in the Asian cohort, but not in European cohort (Figure 3.17,Figure 3.19). The difference in survival indicates fundamental differences in the clinical trajectories of the two ethnic groups.

As basal clusters mainly shows difference in proliferation (e.g. G2M targets) and metabolism (e.g. bile acid metabolism) related pathways, I named them as P (proliferation) and M (metabolism), according to major pathway differences between two subtytpes (Figure 3.20, Figure 3.19). When splitting the two cohorts into three subgroups, the proliferation group (P) in Asians and the metabolism group (M) in Europeans further split into two groups namely P1, P2, M1 and M2, with the number of matched subgroups remained at two (Figure 3.19). The P1 subtype in Asians showed up-regulation of EMT, inflammatory response, as well as angiogenesis pathways (Figure 3.20 A, C, Figure 3.21 A, B). However, P2 has up-regulation of unfolded-protein response (UPR) as well as MYC target genes. The differences in M1 and M2 subtype

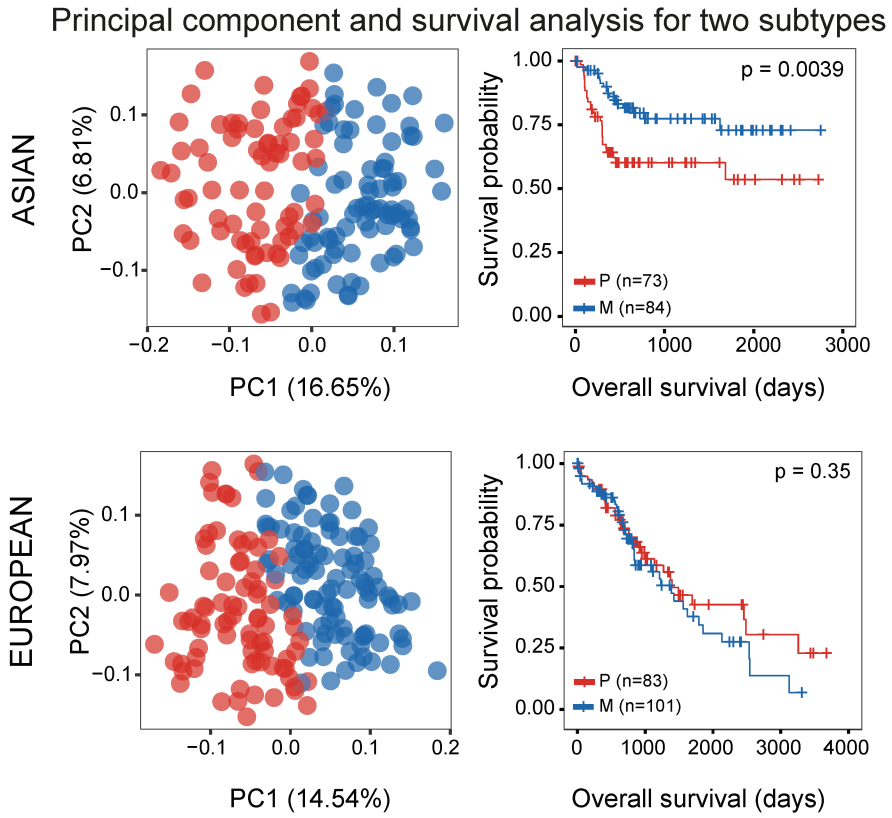


Figure 3.17: Principal components and survival analysis for two subtypes. Both cohorts are well-seperated using 2 subtypes (top-left and bottom-left). While 2 basal split stratifies in the Asian cohort, no separation is observed in the European cohort with two subtypes.

in Europeans overlap significantly with the basal divergence where M1 have up-regulation of cell cycle pathways, but down regulation of metabolic function related pathways such as fatty acid and bile acid metabolisms (Figure 3.20 B, Figure 3.21 B, Figure 3.22).

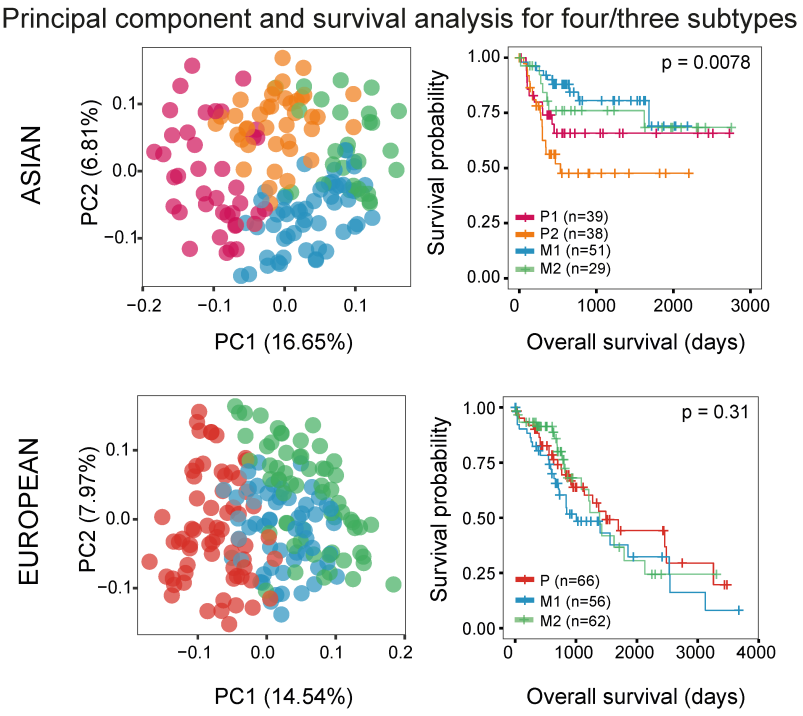


Figure 3.18: Principal components and survival analysis for three and four subtypes. Respective optimum number of subtypes were used for Asian (four) and European (three) cohorts. While subtypes stratify patients in the Asian cohort, no stratification is observed in the European cohorts.

Using solution with the highest cophenetic correlation and consensus silhouette values from NMF output, the next best number of subgroups (two is the first optimal number of subtypes) for Asians and Europeans were found to be 4 and 3 separately (Figure 3.16). When I further partitioned Asians into four subgroups, the good survival group (M) further split into M1/M2, the M1/M2 difference

matches significantly with P1/P2 divergence with M1 having higher expression of immune related pathways as well as EMT (Figure 3.21 A, Figure 3.22).

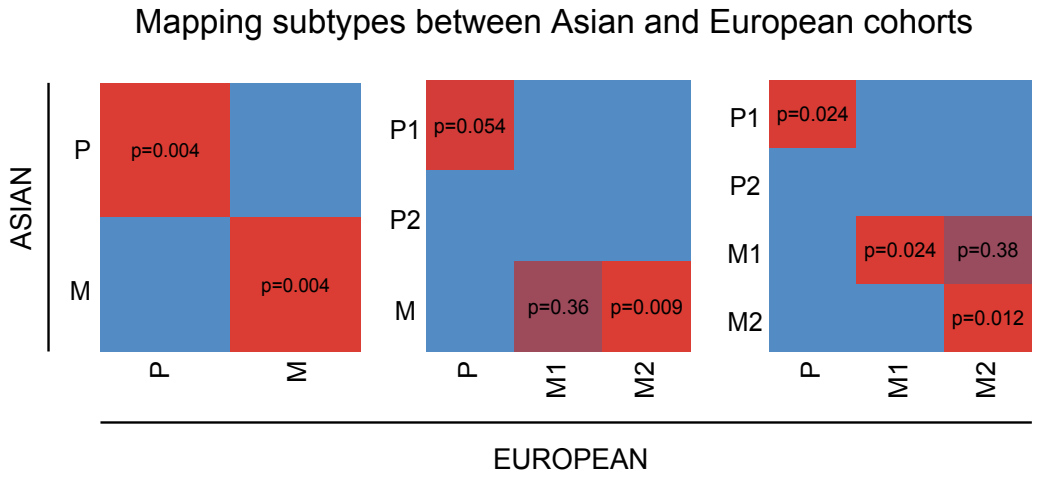


Figure 3.19: Mapping subtypes between Asian and European cohorts using SubMap method. While all three subtypes have a corresponding one in the Asian cohort, P2 subtype is unique to the Asian cohort.

When comparing the four subtypes from Asians and three subtypes from Europeans, the two subgroups within good survival (M1 and M2) match quite well between the two ethnic backgrounds and there is an extra subgroup (P2) unique to Asians (Figure 3.19), indicating important differences between the two ethnic backgrounds. Across all subgroups, RNA subgroups stratify patients very well in Asians, but not in Europeans (Figure 3.17, Figure 3.18).

While three subtypes are shared between Asian and European, there is still differences among these common subtypes. For example, European M1 subtype has up-regulation of cell cycle related pathways compared to M1 (e.g. G2M checkpoint) and they do not show differences in inflammation pathways. However, the Asian M1 and M2 do not show any differences in terms of cell

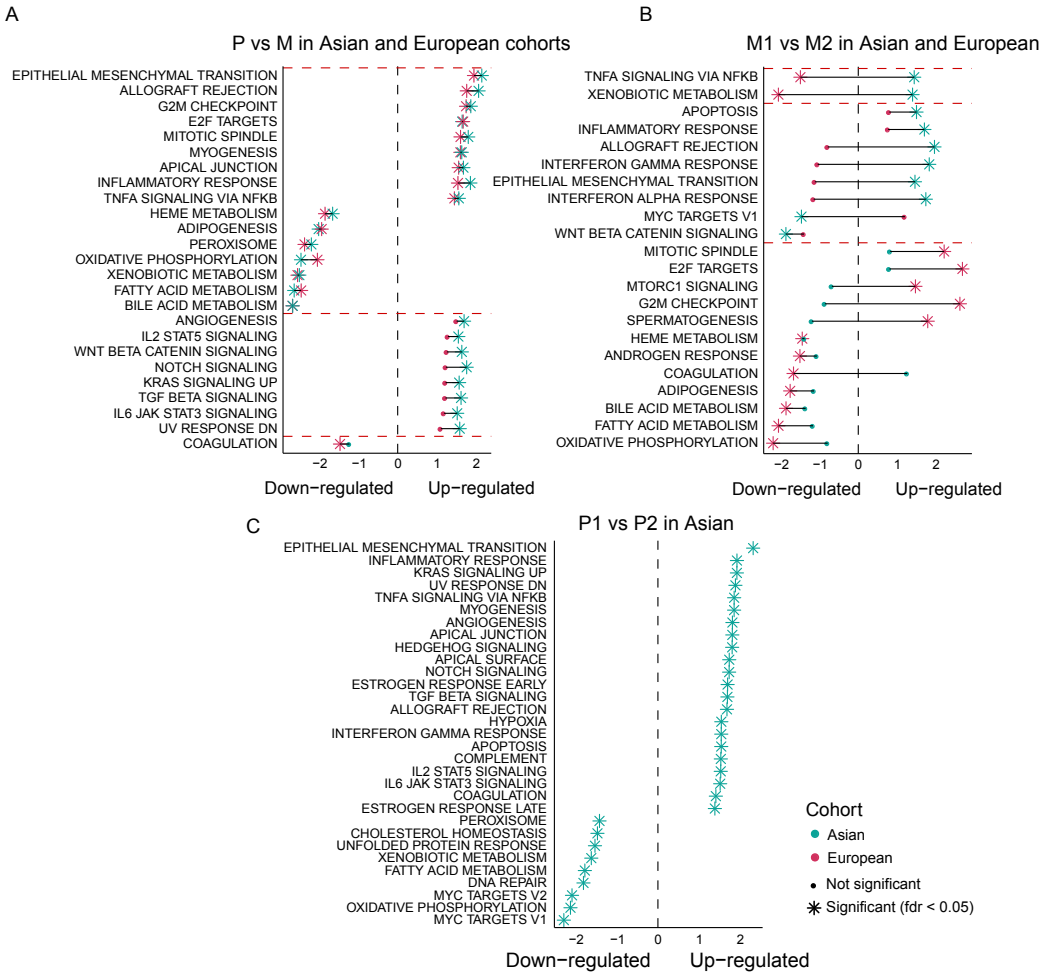


Figure 3.20: Pathway enrichment results for subtypes (A) Comparison of P vs M in both Asian and European cohorts (B) Comparison of M1 vs M2 in both Asian and European cohorts (C) Comparison of P1 vs P2 in Asian cohort. In all panels, x-axis is enrichment score and a positive enrichment score indicates up-regulation while a negative one is down-regulation. Significant enrichment scores are indicated with a star.

cycle related pathways and European M1 is more proliferative compared to European M2 (Figure 3.20 B, Figure 3.22).

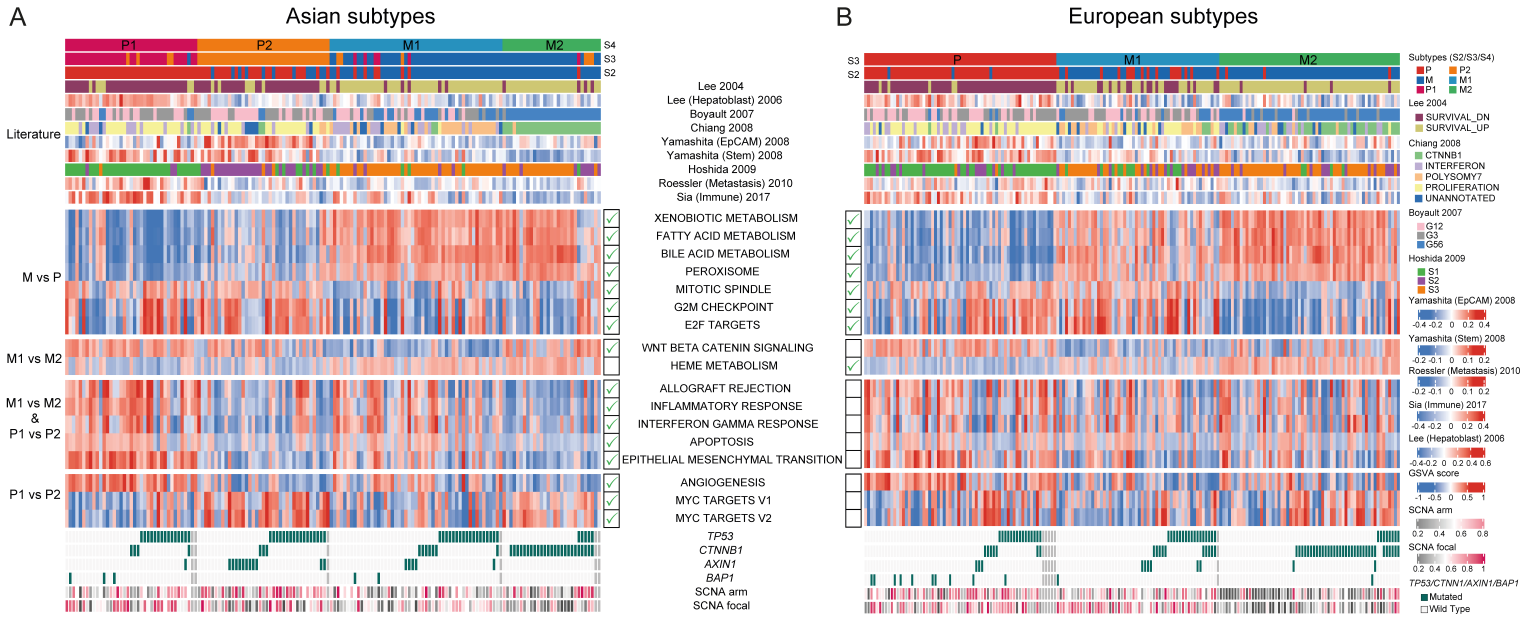


Figure 3.21: Heatmaps showing differentially expressed pathways (A) for Asian and (B) European cohorts. Top rows of each heatmap shows mapping of subtypes to literature subtypes from nine HCC studies. Selected pathways enrichment scores are shown as different blocks based on comparisons (e.g. M vs P). Green tick marks are used to show significant deregulation of each pathway in the marked cohort. Bottom rows are annotations for significantly different genomics changes.

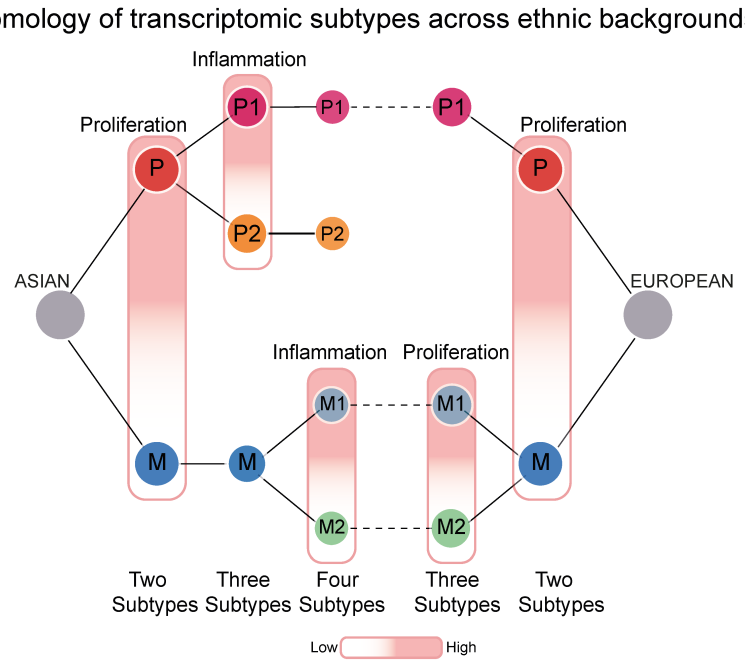


Figure 3.22: Homology of transcriptomic subtypes across Asian and European cohorts. Corresponding subtypes were aligned to show homology between Asian and European HCCs. For each split, the major pathway differences are indicated as a color gradient. For example, in both cohorts proliferation is higher in the subtype B. Inflammatory pathways are major differences at the second and the third splits in Asian cohort while homologous European split is mainly based on the proliferation again.

3.3.5 Clinical and genomic comparison between subtypes

The transcriptomic differences between the subtypes among the two cohorts raised an interesting question: what could have driven the origin of these subtypes. In order to address this question, we can compare the clinical and genomic differences between the subtypes. Some clinical features were showing differences between subtypes in the same direction. For example, both Asian and European subtype P has higher AFP levels and higher number of female and

older patients (Figure 3.23). While no further clinical differences were observed in European subtypes, Asian P vs M and M1 vs M2 showed stage differences where P and M2 are consisting of patients with later stages. In addition to stage, subtype M (good prognosis) included more HBV positive patients in the Asian cohort (Figure 3.23).

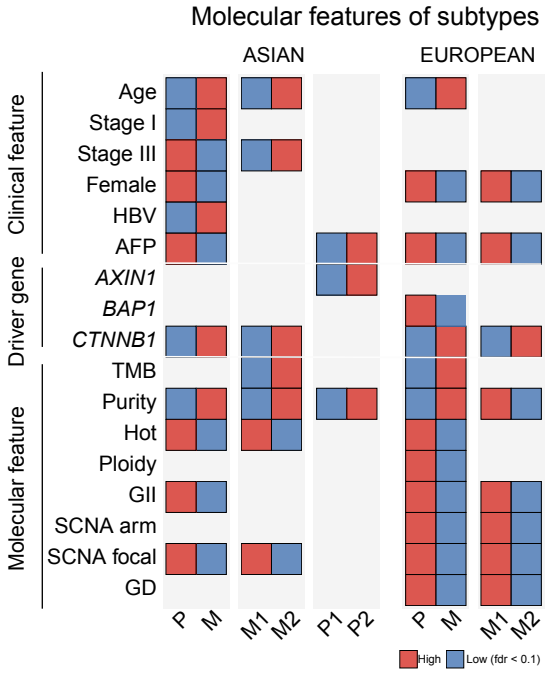


Figure 3.23: Clinical and molecular associations of subtypes. Features from multiple categories were categorized and compared between subtypes P and M and M1 and M2 for Asian (left) and European (right) cohorts. Comparison of P1 and P2 was only conducted for the Asian cohort due the the uniqueness of P2 in this cohort. Fisher's exact test was applied and fdr was calculated using Benjamini-Hochberg method. Assocations with an fdr value < 0.1 are shown.

Looking across molecular events including driver genes, for all the subtypes, I found that molecular differences between the basal split in P and M are highly similar to subsequent partition in M1 and M2 in both Asian and European

cohorts. For example, *CTNNB1* mutation are highly enriched in subgroup M and M2 in both cohorts (Figure 3.21 A-B, Figure 3.23). This indicates extensive connections between genomic events happened in DNA and the transcriptomic changes in RNA. Furthermore, European cohort has higher number of *BAP1* mutations in P subtype compared to M (Figure 3.21 B, Figure 3.23). In Asian cohort, P2 subtype includes higher number of *AXIN1* mutations compared to P1 (Figure 3.21 A, Figure 3.23). In general, stronger genomic correlations in Europeans is obbserved as compared with Asians. For example, both arm and focal level SCNA scores, GII as well as genome doubling (GD) proportions are higher in P and M1 subtypes compared to M and M2 subtypes respectively in the European cohort (Figure 3.23).

3.3.6 A novel subtype driven by genomic changes in the Asian cohort

The Asian specific RNA subtype (P2) is one of the most aggressive subtypes with the highest level of AFP and the poorest survival compared to other subtypes in Asian cohort with a significantly higher level of alpha-feto protein levels (Figure 3.24 A) and poor survival compared to patients from other subtypes (Figure 3.24 B). Existence of such aggressive and ethnic specific subtype raised a series of interesting questions: 1) what are the molecular events specific to this subtype? and more importantly; 2) Do these transcriptomic differences correlate with ethnic differences and explain the origin of this ethnic specific subtype? To address these questions, I systematically compared the genomic features across subtypes.

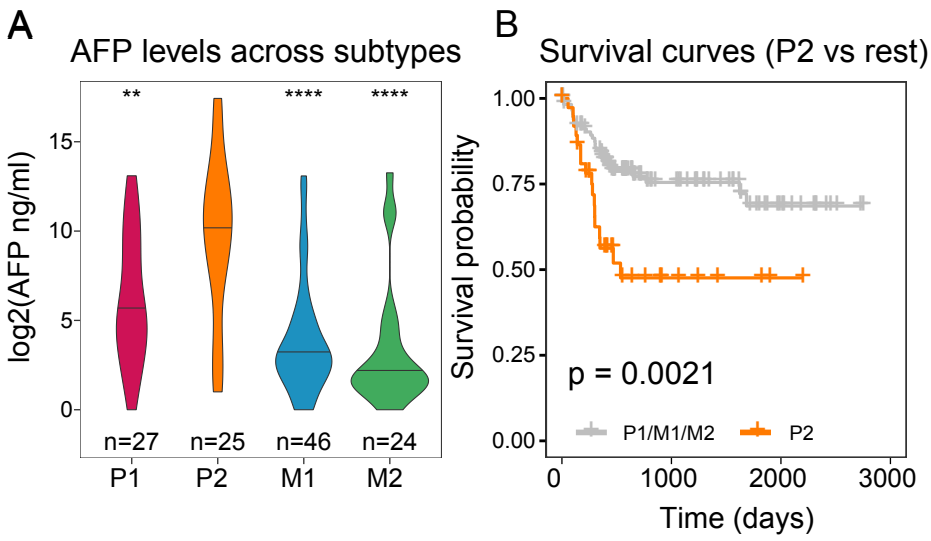


Figure 3.24: (A) Comparison of alpha-fetoprotein (AFP) levels. Horizontal lines at the violin plots marks the median value. When calculating significance P2 was used as the reference group. Wilcoxon's rank sum test was applied. p-values<=0.0001 were labeled as "****", p-values<=0.001 were labeled as "***", p-values <=0.01 were labeled as "**", p-values <=0.05 were labeled as "*" and p>0.05 is "ns". (B) Kaplan-Meier survival curves of P2 vs all other subtypes combined. These findings indicate that P2 is a clinically more aggressive subtype compared to others.

Firstly, comparison of genomic events across subtypes revealed significantly higher frequency of *AXIN1* mutations in P2 subtype (Figure 3.25). Secondly, much higher chromosomal instability was observed in P2 compared to other subtypes using arm level SCNA scores (Figure 3.26 left) from copy number data as well as CIN70 gene expression signature of 70 genes associated with chromosomal instability (Figure 3.26 right) (Birkbak et al., 2011). Thirdly, when we break down the overall copy number comparison to individual chromosome arms, frequency of chromosome 16 deletion is significantly higher in P2 subtype (Figure 3.27). Interestingly, *AXIN1* mutations tend to co-occur with chromosome 16 deletions including both p and q arms rather than individual arm level deletion event (Figure 3.25). Finally, several differentially expressed genes and pathways including overexpression of *AFP* gene (Figure 3.28 A) and MYC targets as well as unfolded protein response (UPR) pathways were observed when comparing to all other subtypes (P1+M1+M2,Figure 3.28 B). UPR is an indicator of endoplasmic reticulum (ER) stress possibly responding to fast cell cycle (Hetz, 2012).

To understand the immune infiltration status of P2 subtype, a recently developed gene signature of was used to calculate an immune score for each sample. This gene signature was obtained by analyzing the expression pattern of immune related cells after deconvoluting tissue expression data into tumor and immune related components (Sia et al., 2017). Based on this signature, P2 is immunologically much colder than the other subtypes with a lower expression of immune class genes with the exception of M1 (Figure 3.29 A). The same pattern was observed, when clustering the patients into hot and cold tumors

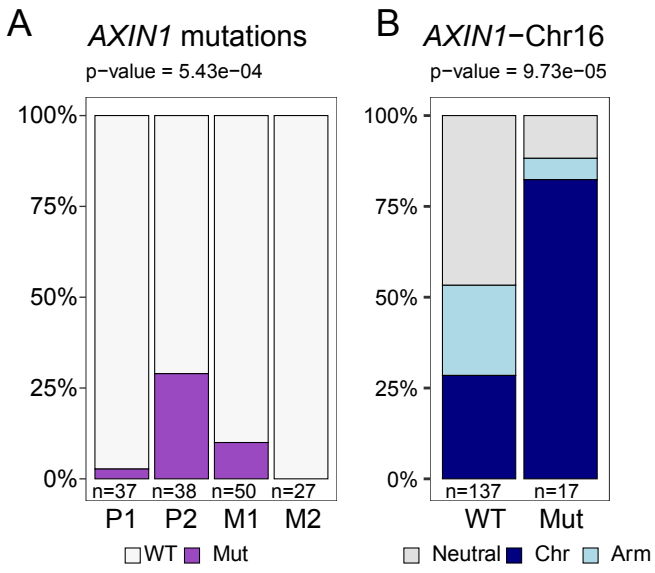


Figure 3.25: (A) AXIN1 mutations across subtypes. P2 subtype has the highest frequency of AXIN1 mutations. (B) Co-occurence of AXIN1 mutations with chromosome 16 deletion. AXIN1 mutations co-occur with the deletion of both arms of the chromosome 16.

based on their overall level of immune infiltration using the deconvoluted score of 14 cell types based on Danaher et al. (2017) (Figure 3.29 B).

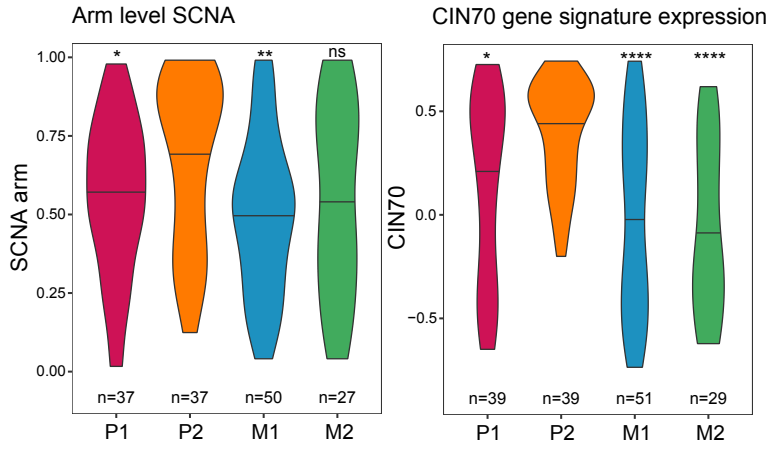


Figure 3.26: Arm level SCNA score and CIN70 gene expression comparison. P2 subtype has higher levels of genetic instability compared to other subtypes. While SCNA score is calculated using genomic data, CIN70 score is a gene expression signature of chromosome instability. Horizontal lines at the violin plots marks the median value. When calculating significance P2 was used as the reference group. Wilcoxon's rank sum test was applied. p-values<=0.0001 were labeled as "****", p-values<=0.001 were labeled as "***", p-values <=0.01 were labeled as "**", p-values <=0.05 were labeled as "*" and p>0.05 is "ns".

Comparing immune components across subtypes, low level of immune infiltration but highest level of myeloid derived suppressor cells (MDSC) are found in P2 subtype. Also, MDSC levels are often much higher in cold tumor compared to hot ones which might be linked to highly immunosuppressive nature of P2 subtype (Figure 3.29 D). In summary, a series of genomic events across multiple layers that are significantly enriched in P2 subtypes is found.

Even though the genomic differences between P2 and other subtypes are quite pervasive, how these events can act concertedly to derive a new RNA subgroup in Asian cohort is quite puzzling. Since ethnic differences are quite minor in

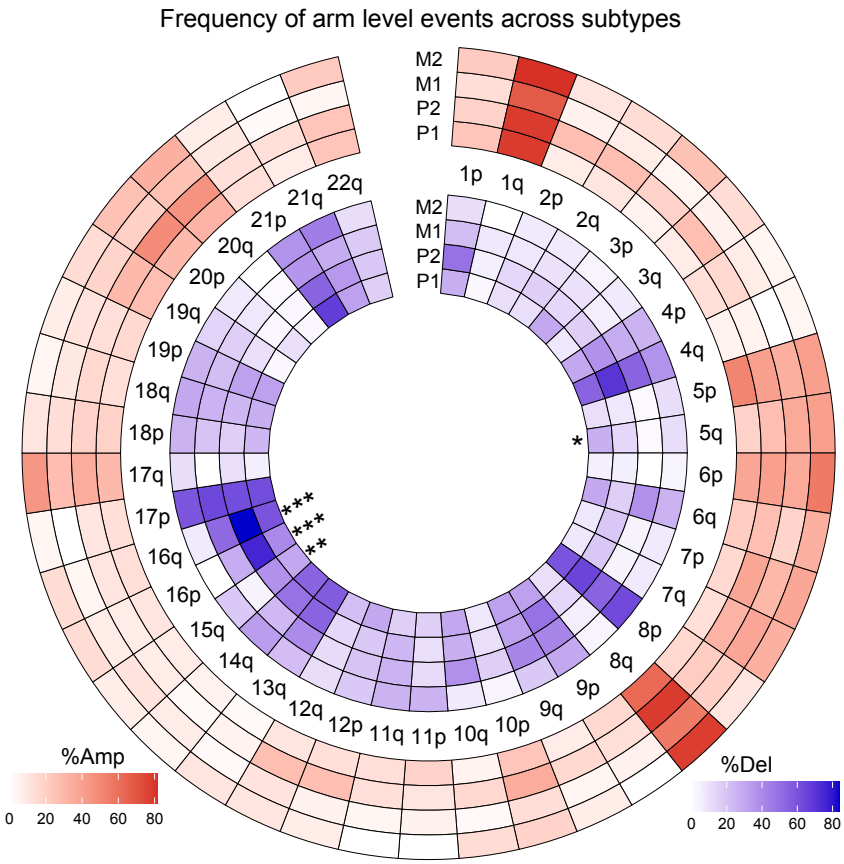


Figure 3.27: Copy number alteration frequency across subtypes. Red color represents amplifications and the blue color represent deletions. Chromosome arm with darker color means the frequency of amplification or deletion is higher for the arm across the cohort. Star sign indicates that frequencies of arm level events across subtypes is significantly different. p-value<=0.0001 were labeled as "****", p-values<=0.001 were labeled as "**" and p-values<0.05 were labeled as "*". Chromosome 16p and 16q arm deletions are significantly higher frequency in the P2 subtype.

driver frequencies, but a lot stronger in copy number alterations I correlated the copy number events and expression levels of all genes across the genome. As expected, most of the CNVs act as cis-regulatory events, positively controlling the expression of genes in the genomic neighborhood (Figure 3.30). Strikingly, CNV at chromosome16, a P2 specific copy number event, tends to impact expression levels of genes across the genome (Figure 3.30). This transcriptomic shift driven by chromosome 16 strongly correlate with the P2 subtype and might explain the origin of this RNA subtype.

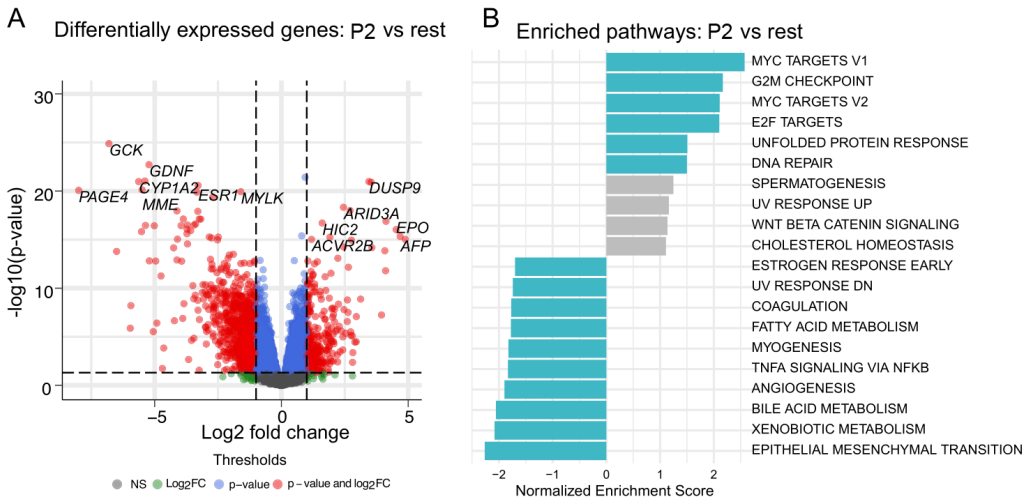


Figure 3.28: Differentially expressed genes and pathways: P2 vs rest (A) Volcano plot showing down-regulated (left red points) and up-regulated (right red points) genes. Top 10 most significant DEGs were labeled. (B) Enriched pathways. Blue color indicates a significant fdr value (q-value < 0.05), gray bars represent non-significant pathways.

In addition to chromosome 16 deletions, other genomic events defining subtype P2 seems to act concertedly with chromosome 16 deletions. For example, previous studies reported a negative correlation between SCNA level and immune infiltration across different cancer types (Davoli et al., 2017). When I

compare the levels of tumor infiltrating lymphocytes (TILs) between low and high SCNA tumors, indeed lower levels of TILs were observed in high SCNA tumors (Figure 3.29 E). Higher genomic instability including chromosome 16 and recruitment of more MDSCs (Figure 3.29 E, F) seems to lead to low immune infiltration in P2.

When I draw a correlation network between multiple events across layers ranging from clinical features, genomic changes, transcriptomic and immune phenotypes, a well-connected network spanning multiple layers that are defining the P2 subtype is observed (Figure 3.31). Taken together, ethnic differences in genome instability seem to trigger a collection of differences defining an Asian specific transcriptomic subtype.

3.3.7 Intratumor heterogeneity (ITH) and integrative survival analysis

Intratumor heterogeneity (ITH) has increasingly been recognized as an important factor driving patient clinical outcome (Mroz et al., 2013; Jianjun Zhang et al., 2014). However, the study of ITH in HCC has been limited to a few multi-sectoring cohorts with small number of patients and limited power in stratifying patients (Zhai et al., 2017). To fully dissect the level of ITH, three different metrics were calculated: 1) percentage of late mutations (pLM), calculated as the fraction of subclonal mutations, 2) Mutant-Allele Tumor Heterogeneity (MATH), measuring the distribution of variant allele frequencies, 3) Shannon's index, calculated based on the subclonal proportions using Pyclone for each tumor. Interestingly, pLM stratify patients when I

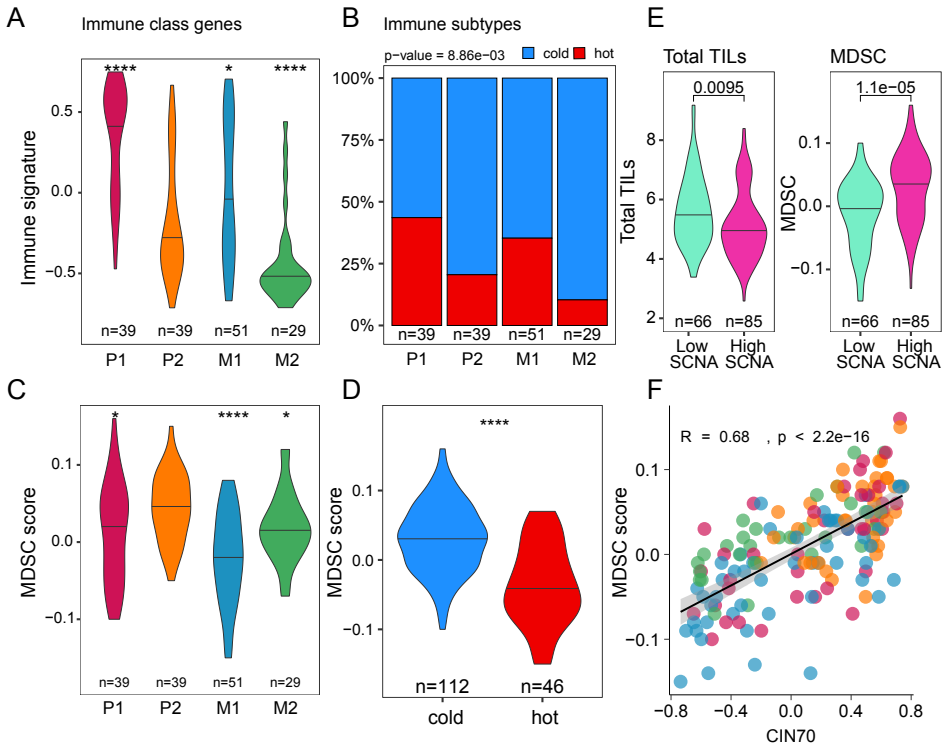


Figure 3.29: Immune related features and P2 subtype. (A) Comparison of immune class gene signature from Sia et al. Horizontal lines at the violin plots marks the median value. When calculating significance P2 was used as the reference group. Wilcoxon's rank sum test was applied. p-values<=0.0001 were labeled as "****", p-values<=0.001 were labeled as "***", p-values <=0.01 were labeled as "**", p-values <=0.05 were labeled as "*" and p>0.05 is "ns". (B) Calculated immune subtypes based on Danaher et al. Similar to immune signature, P2 subtype has lower porportions of hot tumors together with M2 (C) Myeloid drived suppressor cell (MDSC) score across subtypes. (D) Immune subtypes versus MDSC score. Cold tumors have significantly highee MDSC score. (E) Total tumor infiltrating lymphocyte and MDSC level comparison between high and low SCNA tumors. Tumors with high genomic instability have a more suppressed tumor microenvironment. (F) MDSC score versus CIN70 genetic instability gene signature. There is a significant correlation between chromosomal instability and the MDSC score.

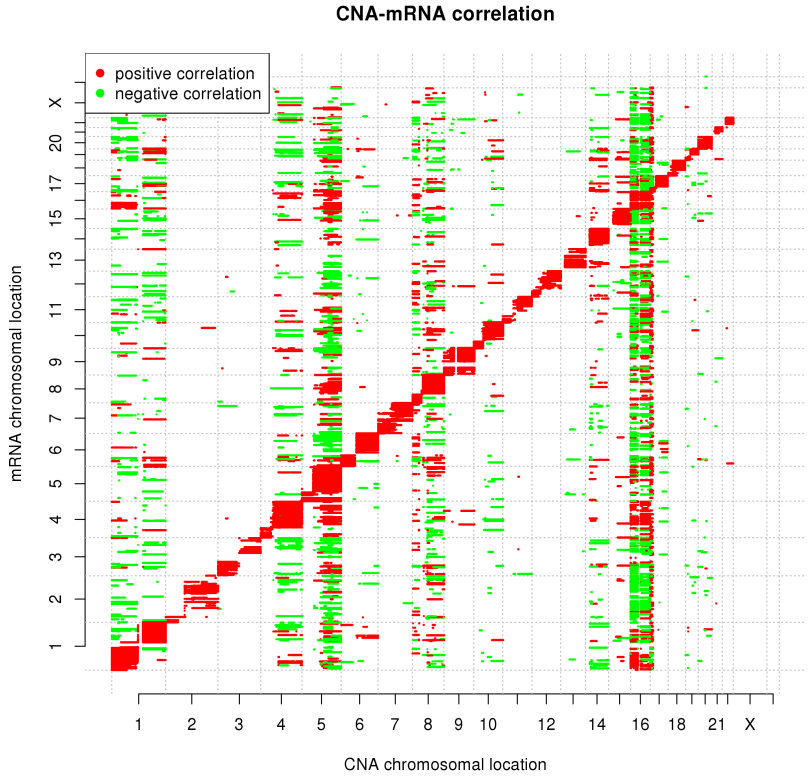


Figure 3.30: Correlation of mRNA expression with copy number across all genes. Spearman's correlation of expression levels and copy numbers (from GISTIC algorithm) of individual genes were calculated across all patients. Red color represent a significant positive correlation meaning amplification of the gene will increase the expression of the compared gene while deletion will decrease the expression. On the other hand, green color indicates a significant negative correlation. As expected, copy number and expression of nearby genes are positively correlated (red diagonal). Interestingly, chromosome 16 copy number is significantly correlated with the expression of genes across the genome. This global significant CNV-expression correlation is also observed for some other chromosomes such as chromosomes 4 and 5.

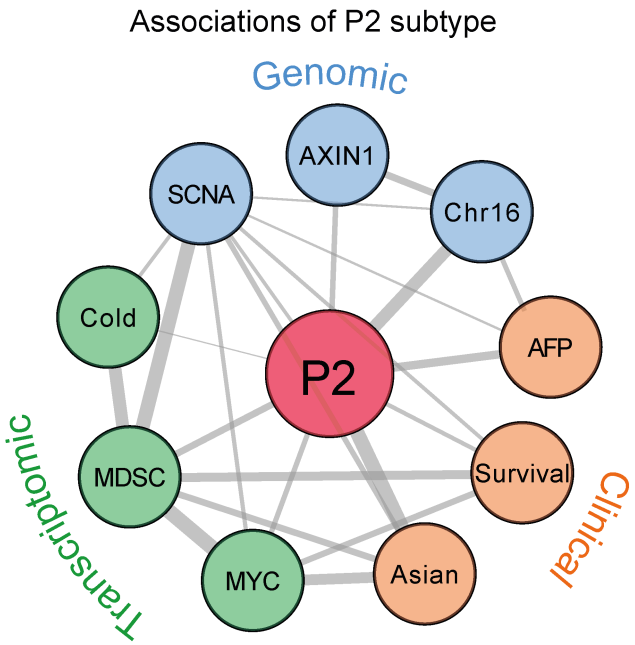


Figure 3.31: Associations of P2 subtype with features from multiple categories. Edges between nodes represents the significance of the correlation. Edges are wider between features that are more significantly correlated. Features are organized under genomic (blue), clinical (orange) and transcriptomic (green) feature categories.

categorize them as low, medium and high levels of each feature (Figure 3.32). When I compared ITH values across Asians and Europeans, two cohorts had similar levels of ITH except for slightly higher pLM in Europeans (Figure 3.33).

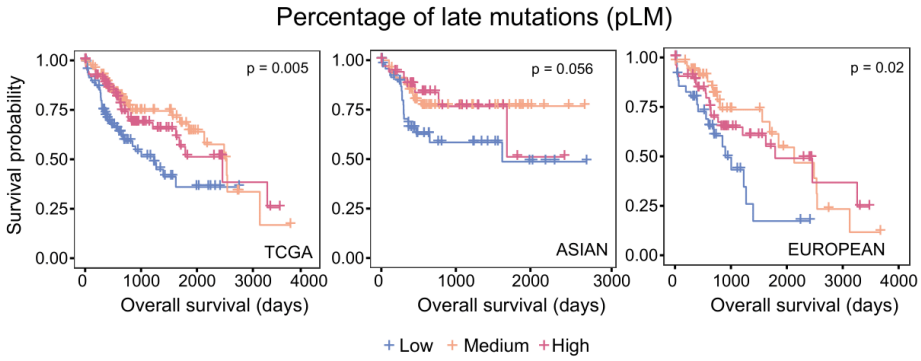


Figure 3.32: Survival curves of categorized percentage of late mutations (pLM). The combined TCGA (left), the Asian (middle) and the European (right) cohorts are shown seperately.

In order to integrate ITH features in a multi-variate model, I first investigate how ITH correlates with other layers and whether ITH can provide independent information in patient survival. I collected 39 features from different layers including clinical features (n=5), molecular features (n=16), drivers (n=12) as well as ITH features as both continuous and categorized versions to account for non-linear relationship with survival (n=6) and only selected features that stratify patient in either of the cohort individually or in the joint cohort for further analysis (log-rank p-value < 0.05, n=19 including 3 clinical, 8 molecular, 6 driver and 2 ITH features).

When I calculate the correlation between features from multiple layers and plot the correlation network for the two cohorts separately, I found that Asian

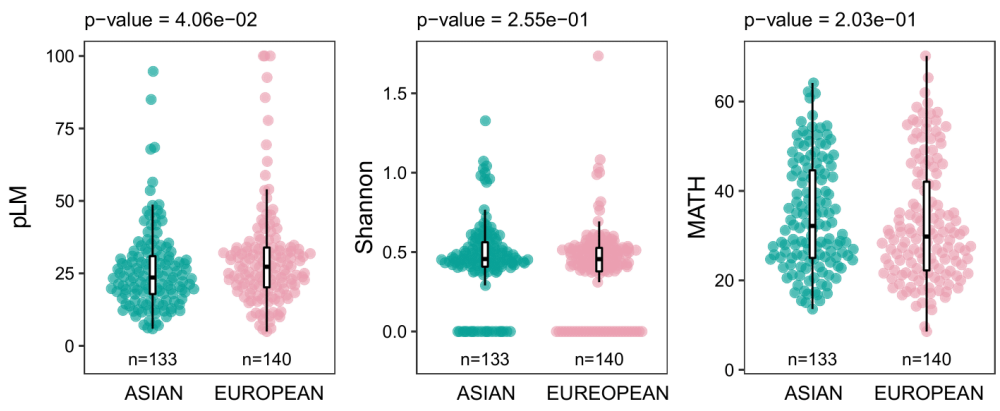


Figure 3.33: Comparison of ITH features across cohorts. p-values are calculated using Wilcoxon's rank sum test. While the percentage of late mutations (pLM) is marginally higher in Europeans (left), Shannon's diversity index (middle) and the mutant allele heterogeneity (MATH) scores (right) were similar between two cohorts.

patients have much stronger correlation in the feature space and ITH metrics tend to correlate with multiple features from other layers. (Figure 3.34). To find the best subset of variables to predict patient survival, I conducted a bi-directional stepwise Cox regression using the combined TCGA cohort and 8 variables were selected for the final model.

Using a likelihood-ratio approach (See Methods), I ranked the importance of variables and found that immune features (e.g. MDSC) and driver genes (e.g. *DOCK2*) are playing very important roles in patient survival (Figure 3.35, see Methods). Interestingly, the importance of variables in the Asian cohort are highly similar to the combined TCGA cohort while European cohort had a very difference profile (Figure 3.35).

Notably, ITH features ranks rather poorly in Asian cohort, but ranked second in European cohort (Figure 3.35, right-bottom). This high ranking of ITH

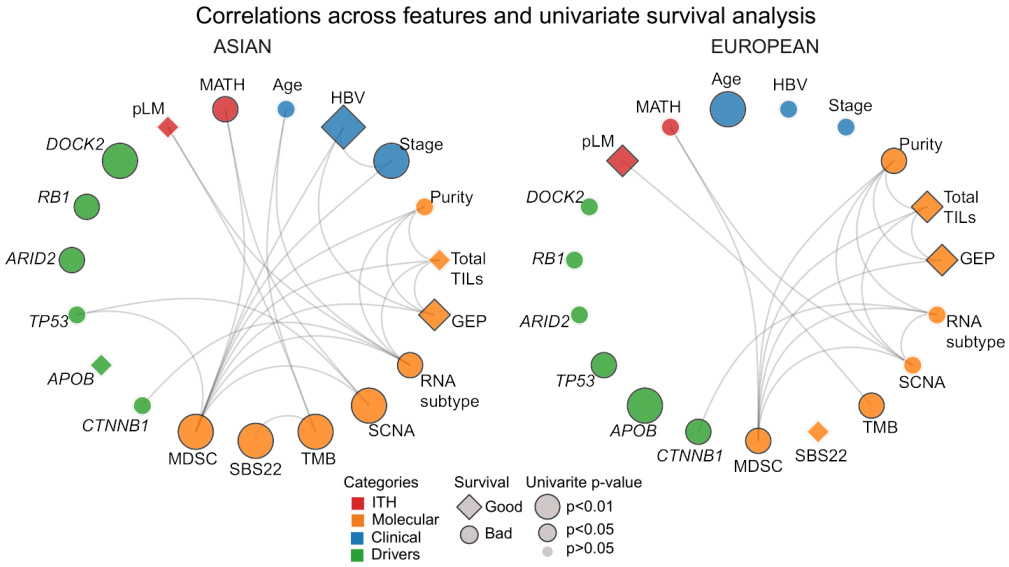


Figure 3.34: Correlation networks for the selected survival variables. Edges of the network indicates correlation between features at the connected nodes. The thickness of edges are re-scaled p-values after -log10(p) transformation. The diamond shape represents a hazard ratio (HR) less than 1 (good prognosis) and circle represents a HR greater than 1 (poor prognosis). For features with multiple levels such as stage, HR of the most significant level is used. The black border around the triangle and bigger size means that feature has a Cox model log-rank score test p-value<0.05 in the univariate analysis. While features with a significant survival prediction are quite different between two cohrots, Asians have more stratifying features.

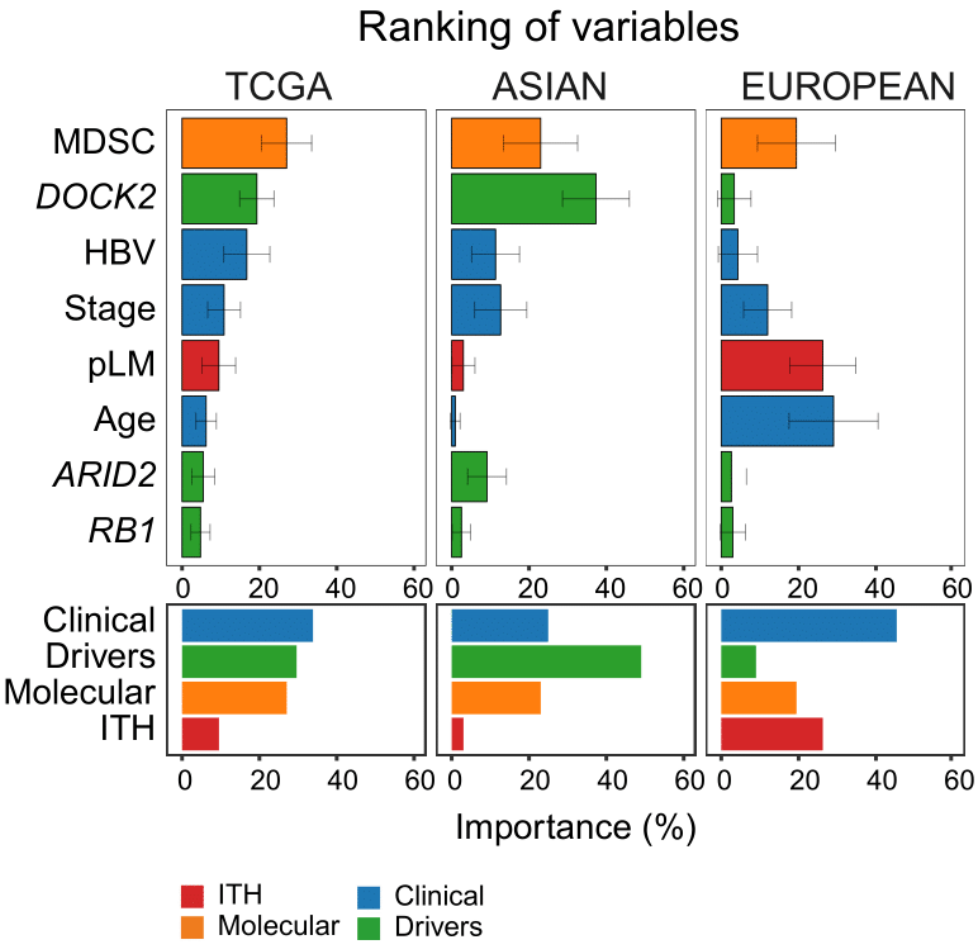


Figure 3.35: The ranking of different clinical, molecular, driver and ITH variables. Chi-squared values from the likelihood ratio test are used as significance indicator. To make robust estimations, importance of each feature was calculated 50 times by randomly subsampling of each cohort. Bars indicate the mean importance across all 50 calculations and the variation of importance scores are shown with an error bar. Ranking of features are quite different between Asian (middle) and European (right) cohorts.

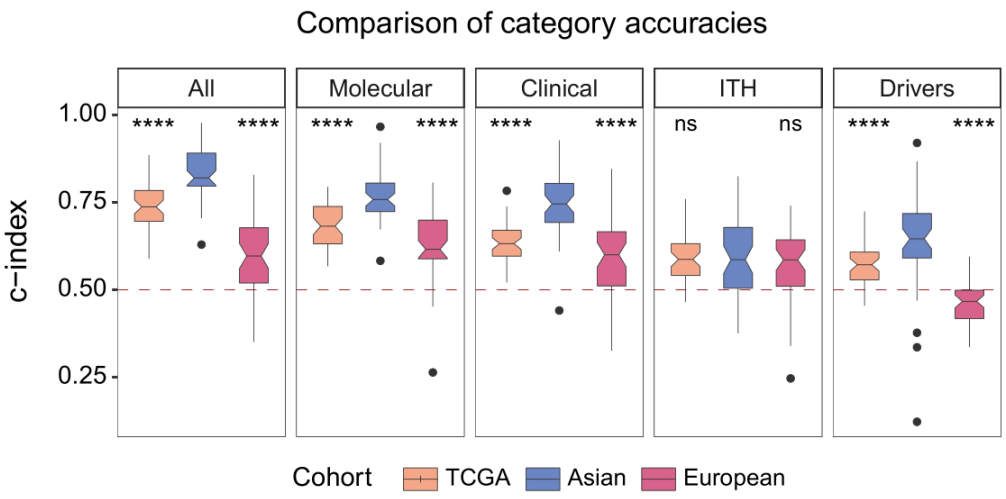


Figure 3.36: The survival prediction accuracy comparison across categeries. Concordance index (c-index) is used as accuracy metric for the comparison. Only a model with a c-index of greater than 0.5 can be acceptable as a good model and the horizantal red dashed line is the indicator for c-index of 0.5. Within each category, Asian cohort was used as the reference group when comparing c-index distributions. p-values<=0.0001 were labeled as "****", p-values<=0.001 were labeled as "***", p-values <=0.01 were labeled as "**", p-values <=0.05 were labeled as "*" and p>0.05 is "ns". Prediction accuracy in the Asian cohort is significantly higher compared to the combined TCGA cohort and the European cohort across all categories except for ITH.

features in the European cohort seem to reflect the poor prognostic ability across all variables in the cohort. When I compared the predictive accuracy between the two cohorts, much higher accuracy is observed (i.e. c-index) in the Asian cohort across categories except for ITH (Figure 3.36). This suggests that patient survival in Asians is more strongly connected to the profile of the tumor and precision medicine will work more efficiently in Asians as compared to Europeans.

3.4 Discussions

3.4.1 Selection criteria for analysis dataset

The reasons for choosing TCGA cohort for ethnic comparison are: Among all available datasets, 1) Access to the raw sequencing data for multi-layer comparison (e.g. Copy number comparison, RNA subtype comparison) is available for TCGA, 2) it consists of balanced number of Asian and European patients unlike other cohorts, and 3) sequencing technologies are uniform which means minimum batch effect.

3.4.2 Rationales for bioinformatics tool selection

Mutational signature analysis: Mutational signature analysis can be categorized under two categoeries: 1) de-novo extraction analysis where all signatures across a cohort is discovered (e.g. BayesNMF by Kasar et al. (2015)) 2) desconstruction of mutations to known signatures (e.g. deconstrucsigs by Rosenthal et al. (2016)). De-novo signature discovery is suitable with whole-genome (WGS) as full collection of mutations is needed to identify the full list of signatures. As TCGA data has mutation information in only coding regions (WES), I used a list of HCC signatures and used deconstruction method by Rosenthal et al. (2016).

Identification of RNA subtypes: RNA subtypes are usually identified by clustering expression data across patients. There are several algorithms for clustering such as hierarchical and k-means clustering methods. Non-negative matrix factorization (NMF) algorithms are also employed for RNA subtype

analysis (Sia et al., 2017). In this chapter Non-negative matrix factorization method is employed as it has at least two advantages compared to classical clustering methods: 1) It does not simply assign samples to individual subtypes but assign weights to each sample for each subtype and 2) together with weights assigned to each sample, genes in the input expression data are also weighted based on their contribution to each subtype. This helps to identify contribution of genes to each subtype and these gene signatures can be used to assign patients in a new cohort to pre-defined subtypes. For example, using gene weights from NMF analysis in this chapter, I assigned a patients in Chapter 4 to TCGA RNA subtypes (see Chapter 4).

3.4.3 Ethnic differences in other cancer types and comparison of ethnic differences in HCC with the literature

Both genetic and environmental agents affect tumor initiation. Accordingly, variations in cancer incidence or outcome across ethnicities in different geographical regions is expected. Understanding these differences in genetic variations and environmental interactions is crucial to stratify patients and develop screening and treatment strategies. One of the most striking ethnic difference with translational implication is significantly higher *EGFR* mutations in Asian lung cancer (40%-64%) compared to white populations (9%-20%) (Tan et al., 2015). Ethnic disparities in somatic genome was also observed for other cancers such as higher frequency of *PTEN* gene deletions in white prostate cancer patients compared to Asian patients (Mao et al., 2010).

Although genomic ethnic comparison of HCC has not been conducted

systematically in the literature, there is a few studies reporting some ethnic differences in HCC. For example, higher proportions of T>A nucleotide change was reported by Totoki et al. (2014). This change potentially corresponds to aristolochic acid signature (SBS22) and the same difference is observed in this thesis with the signature analysis. Morever, an integrated (TCGA, iCluster1) molecular subtype was reported to be dominant in Asian race by Ally et al. (2017). This finding is different from the ethnic specificity of P2 subtype because this iCluster1 also includes patients from European patients with a similar proportions unlike P2 subtype in this thesis.

3.4.4 Comparison of transcriptomic subtypes with the literature

By comparing subtypes identified in the TCGA cohort against the public cohorts, a good concordance between subtypes identified earlier using different ethnic populations is revealed. For example, the basal partition of P and M matches very well with the survival groups found in Lee et al 2004 Lee et al. (2004). Hoshida subtypes S1/S2 and S3 match with P1/P2 and M subgroups in Asians, and S2 (P2) subtype was not observed in Europeans Hoshida et al. (2009). The six subtypes reported in Boyault et al 2007 have a partial overlap in both Asian and European cohorts with G5/G6 subtypes matching well to subtype M in Asians and M2 in Europeans Boyault et al. (2007).

In addition to molecular convergence, features unique to specific ethnic background are also present. For example, hepatoblast subtype found in Lee et al 2004 is mapped to P1 subtype in Asians, but to P subtype in Europeans. Yamashita EpCAM subtype matches to P2 subtype in Asians, but P subtypes

in Europeans. Taken together, mapping the literature subtypes together with the molecular events onto the subtype ontology, both highly concordant and divergent events across the two cohorts were found. Thus, the integrative analysis between the two cohorts identified important ethnic similarities and differences between the two cohorts, providing an important stepping stone integrating information across multiple studies and ethnic backgrounds.

The patient cohort studied in Lee et al., (2004) (Lee et al., 2004) comprises of patients from China and Belgium indicating a cohort with mixed ethnicity. However, they did not conduct any association analysis between ethnic background and molecular subtypes. In their second study, Lee et al used Asian patients as the training set and validated the existence of the HB subtype in European patients (Lee et al., 2006). While this suggests that HB is a common subtype between Asians and Europeans, they did not further analyze the association between ethnicity and subtypes. The patient cohort of Boyault et al. (2007) mainly consist of French patients (80%) which suggest that this is mainly an European cohort. Interestingly, they noted that G1, the subtype with fetal-like expression, mainly comprises of patients from Africa. Even though G1 subtype is a minor subtype that is only ~9% of the whole cohort, the association with African origin suggest a potential ethnic disparity in HCC subtypes. Although Hoshida et al. (2009) identified three global HCC subtypes using mixed ethnicities, they did not discover subtypes using unsupervised techniques within each cohort. Instead, they used a discovery cohort and assigned the rest of the patients to the closest of the three subtypes. This supervised technique might lead to forced assignment of patients to

pre-existing groups. Taken together, highly discordant RNA subgroups seem to correlate with the discovery cohorts from different ethnic backgrounds.

3.4.5 Summary of findings

In this chapter, leveraging the uniform genomic survey across the Asian and European cohorts provided by the TCGA, I have systematically compared the two cohorts across clinical, genomic and transcriptomic features. At the genomic level, higher genomic instability is found in Asians compared to Europeans. Clustering of molecular subtypes revealed a novel Asian specific RNA subtype with much higher genetic instability and immune suppression. Comparison across multiple layers across ethnic backgrounds revealed strong correlation between ethnic specific RNA subtypes and ethnic specific genomic changes. Thus, I observed a strong link between phenotypic differences at the RNA level and genomic changes in the DNA. Lastly, when integrating multiple features across layers including the intra-tumor heterogeneity yield a combined survival model with much better survival prediction accuracy in Asians than Europeans, possibly driven by ethnic specific changes. For the first time, I have provided a comprehensive landscape of ethnic differences in HCC between Asians and Europeans.

3.4.6 Higher genomic instability in Asians

Both inherited genetic factors as well as somatic changes due to environmental exposures can affect tumor initiation. Accordingly, variations in cancer incidence or outcome across ethnicities in different geographical regions is

often expected. Ethnic disparities in somatic genome were often observed for different cancers such as higher frequency of PTEN deletions in white prostate cancer patients compared to Asians (Mao et al., 2010) and higher *EGFR* mutations in Asian lung cancer (40%-64%) compared to European populations (9%-20%) (Tan et al., 2015). In this study, a higher genetic instability for the Asian cohort is observed (Figure 3.3, Figure 3.11). While higher mutation burden in Asians could be explained by external mutagen exposures such as TCM herbs, higher copy number instability is still significant after adjusting for covariates. This could be due to genetic susceptibilities or other unknown causes that remain to be discovered. Understanding these differences in genetic variations and environmental interactions is crucial to stratify patients and develop screening and treatment strategies. The study of ethnic differences in HCC provides novel insights into the origin of HCC.

3.4.7 Complex transcriptomic differences between the two cohorts

Through the comparison of transcriptomic features, I observed a mixture of similarity and dissimilarity across the two ethnic backgrounds (Figure 3.19). For example, at the basal split when partitioning the two cohorts into two groups, a cell cycle active group (P) and a cell-cycle indolent group (M) were found in both cohorts. Thus, the basal split in cancer showed concordance across ethnicities along the proliferation axis (Figure 3.19, Figure 3.22, Figure 3.21). Interestingly, clustering into three subtypes results in the split of M in the European cohort (M1 and M2) and P in Asians (P1 and P2, Figure 3.22). While M1 and M2 also differ in the proliferative features in Europeans, P1 and P2

subgroups in Asians differ in inflammatory features (Figure 3.37). Interestingly, Asian P1 subtype matched to European P but cold subtype P2 is found to be Asian specific. Further clustering of Asian cohort to four subtypes resulted in the split of M. Although these resulting subtypes matched to M1 and M2 in European cohort, major difference in M1 and M2 subtype in Asians was still inflammatory features unlike the proliferation differences between M1 and M2 of the European cohort. Thus, the axis of separation in two ethnic backgrounds are quite different except at the basal split of RNA subgroups.

Moreover, even though the link between immune exclusion and WNT pathway activation has been reported before (Galarreta et al., 2019; Luke et al., 2019) and a subgroup M2 with CTNNB1 mutations was robustly found in both cohorts, overexpression of WNT pathway is only observed in the Asian cohort (Figure 3.37). Furthermore, while Asian M2 is an immunologically cold subtype compared to M1, a reverse relationship is observed between M1 and M2 in European cohort (Figure 3.37 bottom row). Thus, it remains unknown what has yielded this difference in the two cohorts for this subtype. Taken together, the subgrouping in two ethnic backgrounds have a complex landscape of similarity and differences.

3.4.8 High genetic instability drives ethnic specific subtype P2 in Asians

P2 is an Asian specific, aggressive and immunologically cold subtype. I observed multiple genomic events that are linked to P2. Among these, *AXIN* mutations and chromosome 16 deletion significantly co-occurred together and the P2

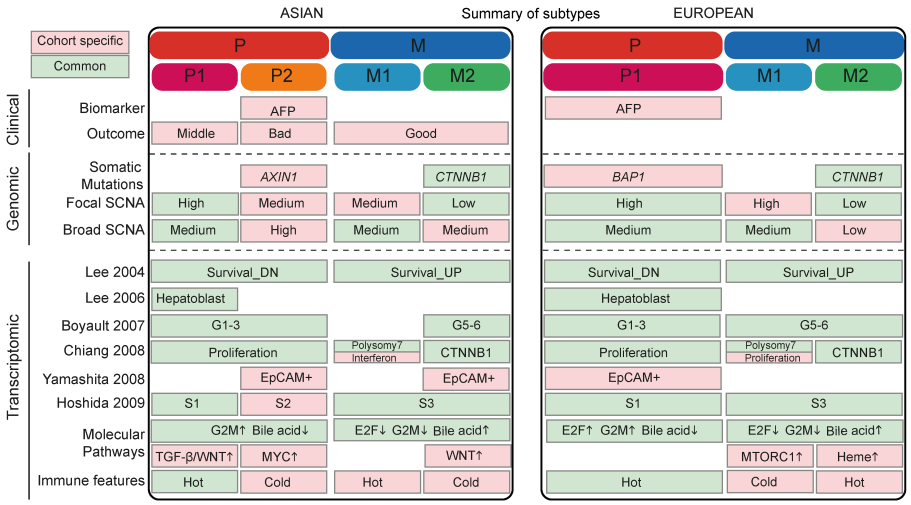


Figure 3.37: Summary of molecular subtypes and comparison of subtype features. This figure is a summary of all findings on transcriptomic subtypes. Similar features between two cohorts are shown with a green color while feature-subtype associations that is linked to only one cohort are shown with red blocks. Corresponding transcriptomic subtypes from the literature is shown at the "transcriptomic" section.

subtype. It is important to note that *AXIN1* gene is also located in chromosome 16. Given the tumor suppressor role of Axin 1 in WNT pathway (Nault et al., 2017; Zucman-Rossi & Nault, 2020), this co-occurrence might be selected in order to inactivate the functional allele of the gene concordant with the two-hit hypothesis (Knudson, 1971).

Higher copy number levels in Asians can be attributable to its unique P2 subtype as it has the highest CNV levels compared to other subtypes (Figure 3.26). Detailed analysis into the copy number differences across subtypes for all chromosome arms also showed link with chromosome 16 (Figure 3.27). Correlation of chromosome 16 copy numbers with gene expression across the whole genome might be the origin of this unique subtype.

Overexpression of unfolded protein response (UPR) pathway in P2 could also be linked to this global transcriptomic rewiring. Because, chromosome 16 is deleted in majority of P2 (>75%) and correlation of its copy number with transcriptome is mostly negative across the genome (Figure 3.30). This might result in global overexpression and protein folding stress which can trigger the UPR activation (Hetz, 2012).

3.4.9 Treatment differences in light of ethnic differences

While immune checkpoint inhibitors (ICI) are starting to gain popularity, predicting patient's response to ICI based on molecular data has become a major challenge for the field (Anagnostou et al., 2020; Grasso et al., 2020). I found that P2 subtype is immunologically cold and has very high levels of myeloid derived suppressor cells (MDSC) (Figure 3.29). Since higher levels

of MDSC levels are also often observed in cold tumors (Figure 3.29), this suggests that MDSCs might be the correlated with the immunosuppressive environment in the P2 subtype. Even though ICI aims to block PD-1, PD-L1 or CTLA4 which are key molecules in T cell suppression, the mechanisms of immune repression by MDSCs can be different which can include the release of immunosuppressive cytokines or arginase (Anani & Shurin, 2017). As a consequence, ICI might not work so efficiently in patients from the P2 subtype. However, studies showed that targeting MDSCs or combining ICI with MDSC targeting therapies might result in better response for tumors with high MDSC levels (De Cicco et al., 2020; Sun et al., 2019; Weber et al., 2018). Thus, the study of ethnic difference might pinpoint new possibilities of novel therapeutic opportunities.

3.4.10 Survival related differences between ethnicities

In addition to genomic and transcriptomic disparities, two cohort show dramatic differences in terms of patient survival. In general, I observed that that more features can stratify for patient survival in the Asian cohort compared to Europeans. One important factor is that Asian cohort has a private molecular subtype (P2) which contributes to patient stratification more strongly than other subtypes (Figure 3.24). Moreover, stratifying features in Asians are more correlated with each other based on the correlation analysis (Figure 3.34) and might contribute to the better predictive power. It is possible that poor liver function in Europeans due to excessive alcohol usage or patient age might explain the worse predictive power in the European cohort. Strong

differences in predicting patient survival between ethnic backgrounds suggest a better opportunity for precision medicine strategies in Asians.

4 Chapter 4: Intratumor heterogeneity and tumor evolution in the PLANET cohort

4.1 Introduction

One challenge to effective cancer therapy is intratumor heterogeneity (ITH) where multiple subclones can co-exist in a tumor resulting in a therapeutic failure as well as acquired resistance due to clonal evolution within the tumor (McGranahan & Swanton, 2017). Several recent studies have reported variable levels of ITH for HCC. However, these studies mostly consisted of small retrospective cohorts. Besides, these studies were often limited to genomic interrogation at the genotypic level (Xue et al., 2016) and phenotypic (e.g. transcriptomic) heterogeneity is underexplored in these studies (Friemel et al., 2015; Ling et al., 2015; Shi et al., 2016; Xue et al., 2016; Zhai et al., 2017). Friemel et al. (2015) have investigated the morphological and genetic ITH in HCC using the multi-regional approach restricting mutational survey to only well-known HCC drivers *TP53* and *CTNNB1* and observed mutational heterogeneity in 5 tumors out of 8 tumors (Friemel et al., 2015). Xue et al. (2016) sequenced multiple sectors from 10 HBV+ HCCs and reported considerable genomic ITH in 4 of these tumors (Xue et al., 2016). Recently, Zhai et al. (2017) and Lin et al. (2017) have applied a multi-sectoring approach for HCC (Lin et al., 2017; Zhai et al., 2017). Lin et al. (2017) sequenced 69 regions from 11 HCC tumors and found that 29% of drivers were heterogeneous among sectors. Zhai et al. (2017) used multi-regional WES and

WGS techniques to study ITH in 9 HCCs as well as metastatic tumors from 2 patients. They observed a variable level of genomic ITH between patients from low ITH to high ITH. These initial attempts to interrogate the intra-tumor heterogeneity levels in HCC have confirmed that HCC shows variable levels of ITH across patients. However, these studies often have low cohort sizes. A comprehensive understanding of tumor heterogeneity and clonal evolution in HCC tumors is needed.

The Precision medicine in Liver cancer across an Asia-pacific NETwork (PLANET) funded by Singapore National Medical Research Council is a prospective study that samples resected HCC from multi-ethnic sites within the established Asia-Pacific Hepatocellular Carcinoma (AHCC) Trials Group. PLANET aims to enroll 100 HCC patients through 6 multi-center trials. When patients were diagnosed with HCC, patient tumors were surgically removed and multi-regional sampling was conducted. After surgical operation, patients were followed up with CT scans every three months. When the patient tumor recurs, the recurrence tumor is also harvested. PLANET aims to capture the natural history of all the patients and understand how ITH might affect the clinical trajectories of HCC patients. In this Chapter, I reported on the initial findings from whole-genome sequencing (WGS) as well as RNA sequencing of the first 67 patients. As I only carried out part of the analysis, I will only report on the findings from these explorations using primary tumors. Also, I conducted an integrative analysis of features that are predicting the recurrence in the PLANET cohort using recurrence-free survival data. As introduce

4.2 Materials and Methods

4.2.1 Patient recruitment and grid sampling

67 patients were recruited under the Translational and Clinical Research (TCR) Flagship Programme named **P**recision medicine in **L**iver cancer across an **A**sia-pacific **NET**work (PLANET) funded by Singapore National Medical Research Council.Liver TCR programme involves three sites in Singapore including National Cancer Centre Singapore, Singapore General Hospital, and National University Hospital, University Malaya Medical Centre from Malaysia, National Cancer Institute from Thailand and Medical City from Philippines. Informed consent was taken from each patient to be collected in the study.

4.2.2 Tissue Extraction and Library Preparation

Total RNA and genomic DNA were extracted from paired normal and tumor tissues using Qiagen AllPrep DNA/RNA Mini Kit. Quality check of the extracted DNA and RNA were conducted with gel eletrophoresis and Agilent 2100 Bioanalyzer respectively. For WGS samples, qualified genomic DNA were shortened by sonication using Covaris system and quality check of the fragments were performed with Agilent 2100 Bioanalyzer. The fragments were end-repaired, adaptor-ligated, amplified and sequenced by Novogene-AIT sequencing company. mRNA libraries were prepared and sequenced by the sequencing platform at Genome Institute of Singapore. Tissue extraction is done by our former colleague in the Genome Institute of Singapore Jia Qi Lim.

4.2.3 Somatic mutation calling

Raw paired end read were mapped to human reference genome (GRCh37) using Burrows-Wheeler Aligner (BWA) (version 0.7.12) (H. Li & Durbin, 2010). Duplicated reads were removed using sambamba (version 0.6.4) (Tarasov et al., 2015) and base quality recalibration and local realignment were conducted using the Genome Analysis Toolkit (GATK, version 3.1-1). Somatic point mutations were called using Mutect (1.1.7) algorithm by comparing normal and tumor samples (Cibulskis et al., 2013). Strelka (version 1.0.14) was used for indel calling (Saunders et al., 2012). Somatic mutation calling for the PLANET cohort was carried out by Dr. Hechuan Yang in the Genome Institute of Singapore.

4.2.4 Driver gene identification

Details of driver gene identification is described in Chapter 2. In short, somatic mutation data of five big HCC cohorts were complied. Somatic mutations were annotated using Oncotator (Version 1.9.2.0). Three different driver identification algorithms including MutSigCV (Lawrence et al., 2013) (a method based on mutation rate), TUSON Explorer (Davoli et al., 2013) (a method based on clustering of functionally important mutations) and 20/20+ (Tokheim et al., 2016) (a machine learning approach unifying multiple features of somatic mutations) were used. In total, 62 candidate driver genes were identified by combining drivers (q-value <0.1) from three algorithms.

4.2.5 De novo signature analysis and timing of signatures

de novo mutation signatures were inferred using BayesNMF (Kasar et al., 2015). Optimal number of signatures was found to be 18 based on 50 independent iterations of NMF and mutations were decomposed to these *de novo* signatures using deconstructsigs (Rosenthal et al., 2016). After filtering out the low frequency signatures (those with mean proportion across samples < 0.02 or maximum proportion across the cohort < 0.2), 8 de novo mutational signatures were maintained These signatures were found to be highly correlated to 15 known COSMIC signatures (cosine similarity >0.7) (PCAWG Mutational Signatures Working Group et al., 2020). Mutations were subsequently decomposed to these 15 known signatures using deconstrucsigs and 9 final mutational signatures remained after filtering signatures with low contributions (mean proportion across samples < 0.02 or maximum proportion across the cohort < 0.2).

In order to understand the history of these the mutational process, I partitioned the mutations into truncal (shared across all sectors) and non-truncal events. Signature contributions were inferred separately for the truncal and non-truncal mutations. Comparison of truncal versus non-truncal mutation proportions were conducted using paired Wilcoxon signed-rank test.

4.2.6 Arm level copy number alterations

Somatic copy number variations were identified using the Sequenza (version 2.1.2) (Favero et al., 2015) and segmental copy numbers were converted to

arm and cytoband level CNVs using the GISTIC algorithm (Mermel et al., 2011). In order to compare CNVs in our cohort with the public cohort, GISTIC was used to find significantly amplified or deleted regions in the TCGA cohort (TCGA-LIHC). Genome Wide 6 SNP array Level 3 data (copy number segmental information) was downloaded from the GDC Data portal. In order to identify arm level events, ploidy adjusted copy number of segments were calculated by subtracting ploidy from total copy number. Then, cytoband annotation were performed using position information. For each chromosome arm, segment lengths were summed if they have copy number event in the same direction. Chromosome arms were labeled as amplified or deleted if at least 70% of the arm has alteration in the same direction. If a chromosome arm is amplified and deleted across all sectors, it was denoted as truncal, otherwise it is denoted as non-truncal event.

4.2.7 Focal copy number analysis

To find subclonal and clonal focal CNVs, significantly altered cytobands (q value<0.01) were first identified from the TCGA cohort. Subsequently, amplification and deletion events were identified in our samples using thresholded copy number output from the GISTIC results. If all samples have same CNV event for a cytoband, it was denoted as a clonal CNV. If the CNV event occurs to a subset of sectors, this event will be called as a subclonal event. In order to annotate important driver genes in the cytobands, combined genes from the Cancer Gene Census genes (CGC, n=723), liver driver genes (n=62) and a subset of driver genes that were reported in a few large-scale

HCC studies (Shibata & Aburatani, 2014). Gene labels are shown if the gene has oncogene role based on cancer gene census list and amplified or tumor suppressor and deleted.

4.2.8 Comparison of event frequency propotions between truncal and non-truncal events

To compare high and low frequency event proportion comparison between truncal and non-truncal events in drivers, genes were labeled as high if mutational frequency is greater than 10% across the combined HCC cohort complied in Chapter 2 (n=1349) and labeled as low frequency if the frequency is less than or equal to 10%. Then, Fisher's Exact test was applied for the comparison of truncal and non-truncal mutation proportions in high and low frequency driver genes. Similarly, focal copy number events in the TCGA cohort were used to compare truncal and non-truncal event propotions in high and low frequency CNV events. For CNV event comparison, median frequency of peak event was used as threshold to dichotomize CNV events as high and low CNVs.

4.2.9 Calculation of DNA ITH

For DNA ITH, union of all mutations were taken across all sectors and presence/absence matrix was generated using all mutations. Jaccard's similarity index (JI) was calculated and 1-Jaccard's index was used as genetic divergence between each sector pair within a tumor. Formula for calculation is shown below:

$$DNA\,ITH = 1 - \frac{x}{x + y + z}$$

In the above formula, x denotes the number of shared mutations between two sectors and y and z are number of private mutations in each sector. The mean value of the dissimilarity between all possible sector pairs within a tumor is used as the DNA ITH.

4.2.10 Identification of RNA subtypes

Raw RNAseq reads were aligned to human genome (hg38) using STAR (Dobin et al., 2012) and read counts for transcripts were obtained using featureCounts (Liao et al., 2014). Only protein coding genes were kept for further analysis. Raw reads for coding genes were normalized using DESeq2 and log2 of 1 pseudo count added values were obtained (Love et al., 2014). Top 3000 most variable genes were selected using the median absolute deviation. Non-negative matrix factorization was used to identify RNA subtypes (Gaujoux & Seoighe, 2010). 200 runs of NMF was conducted with Brunet's algorithm. NMF algorithm deconvolutes count matrix as multiplication of two matrix. One of these has ranks as columns and weights/loadings for each feature (gene) as rows. In the other matrix, sectors/samples are given in the columns and weight for each rank/subtype is given as rows. Each sector was assigned to the NMF rank with the highest weight based on the second matrix described. To assign new subtypes to TCGA Asian subtypes, gene weight matrix of two cohort used.

Overlapping genes were used and Spearman's correlation of gene weights were calculated for all rank combinations. RNA subtypes were assigned to the name of the TCGA-Asian subtype with the highest correlation.

4.2.11 Calculation of RNA and immune ITH

Normalized counts of all protein coding genes were used to calculate RNA ITH. Using these expression levels, Spearman's correlation coefficient was calculated between all sector pairs and 1-Spearman's correlation was used as transcriptomic divergence metric. Similar to DNA ITH, the mean value of the dissimilarity (1-Spearman's correlation) between all possible sector pairs within a tumor is used as the RNA ITH.

Immune subtypes were calculated as described in Chapter 3 based on a previous study (Danaher et al., 2017). In summary, 14 immune cel type score were inferred and total tumor infiltrating lymphocyte scores (TILs) were calculated as the mean score of all cell types which correlates to CD45 cells with a Pearson's correlation coefficient of greater than 0.6. Using the scores of 14 cell types as well as TIL scores (n=15), all sectors were clustered using k-means algorithm (k=2) and cluster with higher expression of these immune cell types are denoted as "hot" and "cold" otherwise. If a tumor has sectors in both hot and cold subtype, this patient was denoted as "mixed" for immune subtype. Immune ITH was determined by calculating 1-Spearman's correlation between all sector pairs using Danaher cell scores (n=15) and the mean value of the dissimilarity (1-Spearman's correlation) between all possible sector pairs within a tumor is used as the immune ITH.

4.2.12 Feature correlation and Integrative survival analysis

From the patient cohort, several important clinical, molecular as well ITH features were collected. These included important clinical features (n=4, stage, sex, age, viral status), molecular features (n=8, AA signature, aflatoxin signature, mutation status of driver mutations at TP53, CTNNB1 as well as TERT, RNA subtypes, genome doubling, immune subtypes) as well as ITH features (n=3, DNA, RNA and immune ITH). In order to calculate the correlation, Fisher's exact test was used in case of two categorical variables. When both variables are continuous, linear regression was used to calculate the correlation between them. In the case of mixed categorical and continuous variables, Kruskal–Wallis analysis of variance test was applied.

For a perspective cohort, I used the patient relapse free survival integrating all 15 variables in the model (using the coxph function in R). In order to compare variable importance, likelihood ratio test was applied to the model with and without each variable across all variables. Likelihood ratio chi squared value were used as the indicator of importance using Anova function in R. Importance is calculated by summing all the chi squared values and calculating the percentage of each variable. Using the multivariate cox model, I divided patients to three groups based on the predicted hazard values. Kaplan-Meier survival curves of three survival groups were plotted using the surv_fit function in R.

4.3 Results

4.3.1 Patient cohort and sequencing

In this study, 67 HCC patients with different ethnicities including 46 Chinese, 7 Malay, 4 Thai, 5 Indonesian, 3 Burmese, 1 Cambodian and 1 Indian were recruited under the PLANET study. While majority of patients had solitary HCC (77%), there were multi-focal cases (23%) as well. Majority of patients are HBV positive (n=40), three patients are HCV positive and 23 patients are non-viral cases. The cohort consisted mainly of male patients (72%) and the median age of patients is 67.

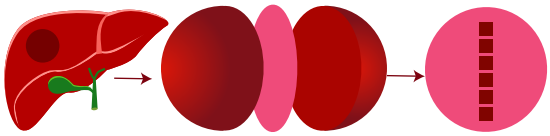


Figure 4.1: Schematic representation of grid sampling method. A central slice is taken out from the tumor and consecutive sectors were sampled along a grid line. Number of sectors were determined based on tumor size.

In order to survey intratumor heterogeneity (ITH), multi-region sequencing was applied to all tumors by taking a central slice from the tumor and sampling consecutive biopsies across a grid (Figure 4.1). In total, 331 samples were sequenced using both whole genome sequencing (WGS, n=318) and whole exome sequencing (WES, n=13) technologies with an average depth of 46x for WGS and 85x for WES samples. Among these, 67 samples were from adjacent normal tissues which were sampled in order to identify somatic changes in tumors. In addition to WES and WGS sequencing, mRNA of 253 samples

from 67 patients were sequenced using RNA-seq.

4.3.2 Mutation burden and mutational signatures

By comparing genomes of tumor sectors against the normal genome, somatic mutations were detected among multiple sectors of the same tumor. Across all 67 patients, mutation rates vary greatly ranging from 0.78 to 21.2 mutations/Mb (median of 4.8 mutations/Mb) (Figure 4.5). In Chapter 3, mutational signatures were directly deconvoluted to known literature-reported mutational signatures. With the WGS data, estimation of de-novo signatures can be achived with a much higher accuracy compared to WES data. Hence, I conducted de-novo signature discovery using BayesNMF in PLANET cohort to dissect mutational processes in the history of HCC tumorigenesis. As a result, 8 novel mutational signatures are identified. Interestingly, these 8 de-novo signatures strongly correlate with 15 known signatures (Figure 4.2 A) from the extended COSMIC signatures with a cosine similarity greater than 0.70 (Forbes et al., 2017; PCAWG Mutational Signatures Working Group et al., 2020).

By performing spectrum decomposition and keeping signatures with appreciable proportions (mean proportion $>$ 2% or max proportion $>$20%), I partitioned the mutational landscape into contributions of these 9 COSMIC signatures (Figure 4.2 B, Figure 4.3) (Rosenthal et al., 2016). While classical age-related signatures SBS1 and SBS5 contributed a high proportion of the somatic SNVs (19% and 32%), a new signature SBS40 was also found to be in high frequency in this cohort (7%). In addition, signature SBS4 (smoking, 8%), SBS6

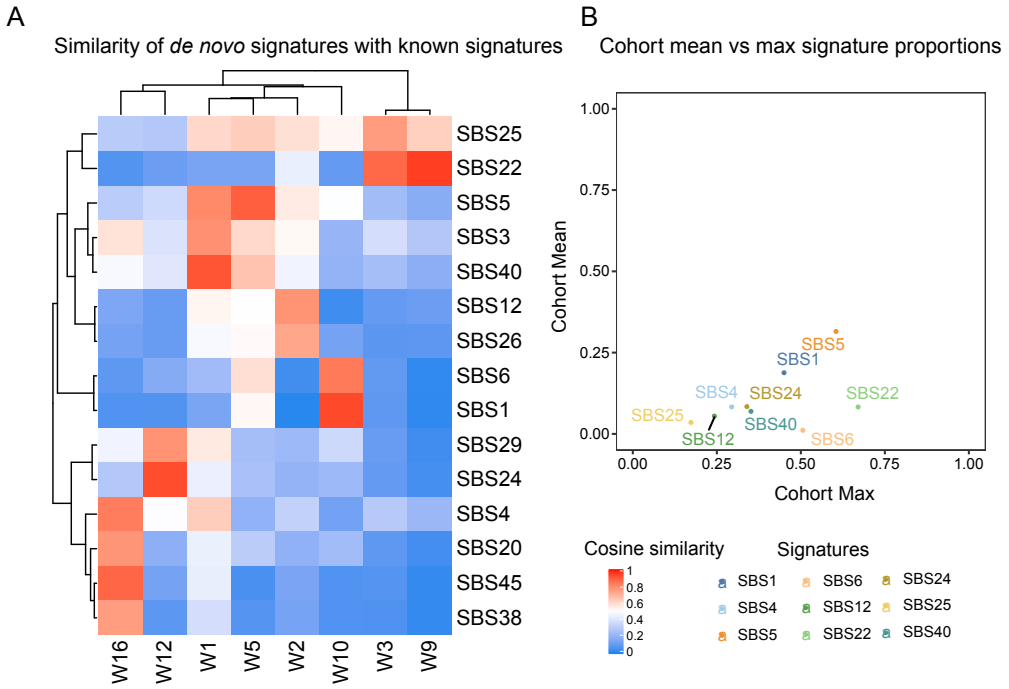


Figure 4.2: Mapping de novo signatures to COSMIC signatures. (A) Cosine similarity of de novo etracted signatures in PLANET cohort to known COSMIC signatures. COSMIC signatures with cosine similarity to novel signstures greater than 0.70 were kept and used for further signature deconvolution. (B) Mean versus maximum proportion of deconstructed signatures across the PLANET cohort. Among 15 COSMIC signatures similar to de-novo extracted ones, 9 signatures with an appreciable proportion in the cohort were kept for further analysis (mean proportion $> 2\%$ or max proportion $> 20\%$).

(DNS-mismatch, 1%), SBS12 (unknown etiology, 6%), SBS22 (aristolochic acid (AA) exposure, 8%) and SBS24 (aflatoxin B1 exposure, 9%) were also detected in the PLANET cohort. It is very interesting to observe that the proportion of AA signatures (SBS22) strongly correlate with the total mutation burden similar to the Asian cohort in Chapter 3 (Figure 4.4).

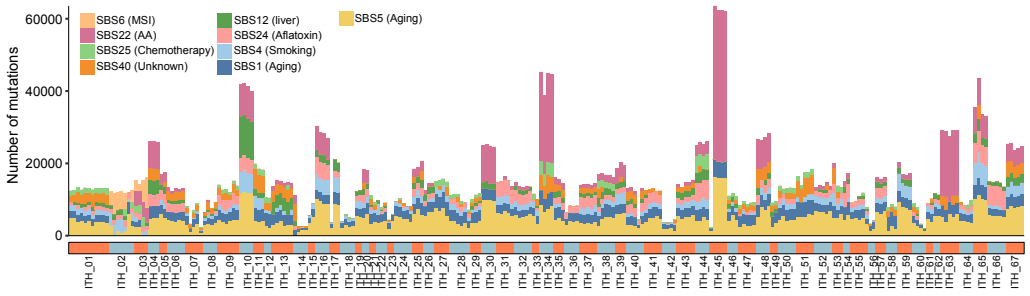


Figure 4.3: Number of mutations attributed to each signature type across samples. It is important to note that samples with high tumor mutation burden have aristolochic acid (AA) signature contributions (pink). Patient annotation is shown below.

4.3.3 Driver genes in the PLANET cohort

Using 62 driver genes identified earlier (see Chapter 2). 47 of these drivers had non-synonymous mutations in the PLANET cohort. *TP53* (49%) and *CTNNB1* (31%) were the most frequently mutated drivers similar to the public cohorts (Figure 4.5). Known HCC drivers such as *ALB* (12%), *ARID1A* (10%) and *ARID2* (9%) were among recurrently mutated genes. Different from samples in Chapter 2, the PLANET cohort was sequenced using the WGS technology. Thus, we are able to identify mutations in non-coding regions. As previously reported by other studies, mutations in the promoter of *TERT* gene is one of the

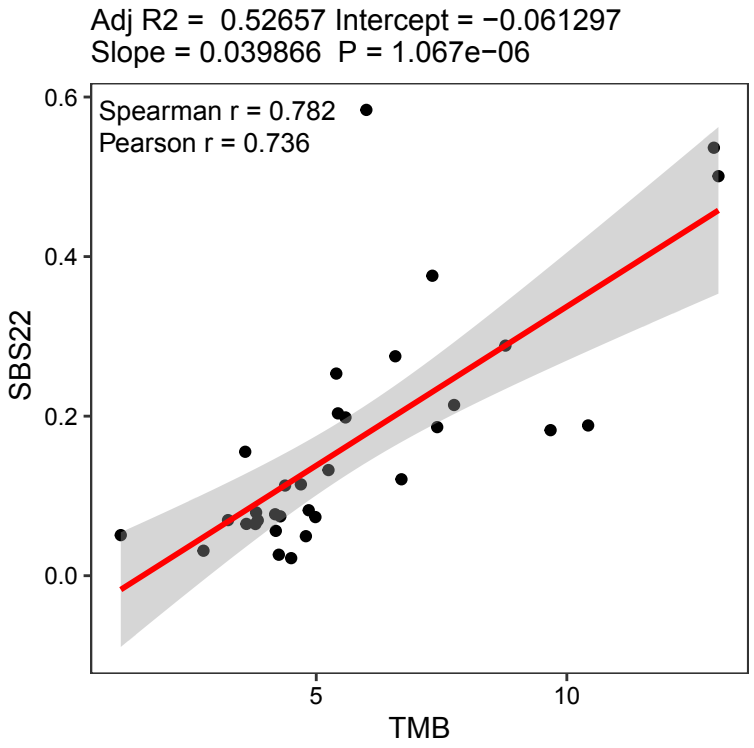


Figure 4.4: Correlation between mutation burden and AA signature proportion in the PLANET cohort.

driver events in HCC (Jhunjhunwala et al., 2014; Shibata & Aburatani, 2014). Classical activating *TERT* promoter mutations G228A (25%) and G250A (3%) were also frequently observed in this cohort (Figure 4.5). In addition to known driver genes, several novel HCC drivers were also found in the PLANET cohort such as *FRG1* (6%) and *ATRX* (4%).

4.3.4 Recurrent drivers are early

Cancer is multi-stage process including tumor initiation, progression and metastasis (Lytle et al., 2018). While driver genes might drive tumor initiation, they can also play important roles in tumor progression. With the multi-region sequencing technology, timing of drivers can be determined by looking at presence of mutations across the tumor sectors. To understand the timing of drivers in the history of tumorigenesis, mutation status was evaluated for each driver across the sectors. It is interesting to observe that most of the identified driver mutations are truncal events and tend to be shared across sectors (Figure 4.6).

Moreover, drivers that are less frequent at the population level across patients also tend to happen at the subclonal level within a patient (p-value=0.009, Figure 4.7 left). For instance, 79% of *TP53* mutations and 57% of *CTNNB1* mutations were truncal (Figure 4.6). On the other hand, the rare novel driver gene *FRG1* is non-truncal in all the mutated cases (n=5) likely indicating a driver of tumor progression. This suggests that there are potentially many subclonal driver genes in HCC, which might be missed by the conventional single sector sequencing approach, but could be discovered by our multi-region

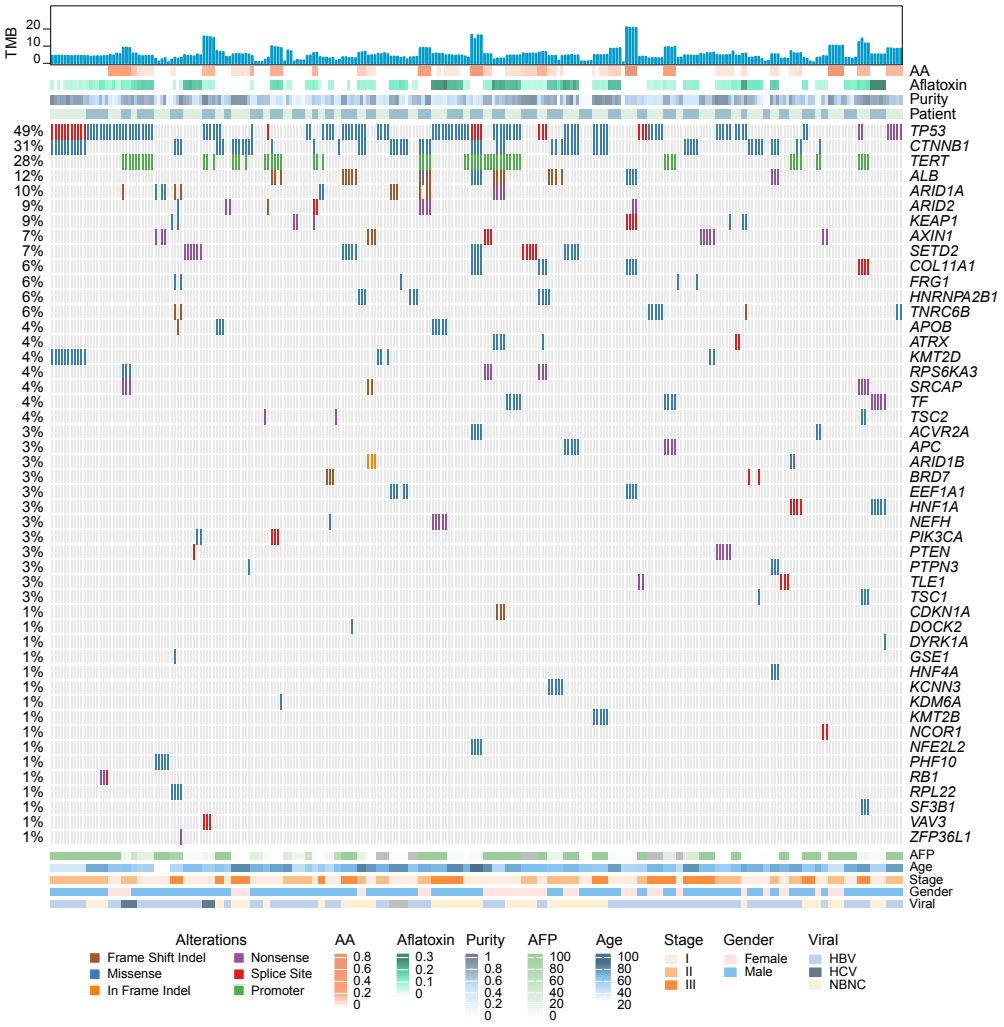


Figure 4.5: Drivers in the PLANET cohort. Each column is a patient and each row is a gene. Type of mutations were indicated with different colors. Percentage of mutations across cohort is shown at the left side (percentage is calculated on patient level, not sample). Multiple sectors belonging to one patient were annotated (Patient annotation). Mutation burden is plotted as a bar plot on the top. Clinical features were shown at the bottom annotation panel.

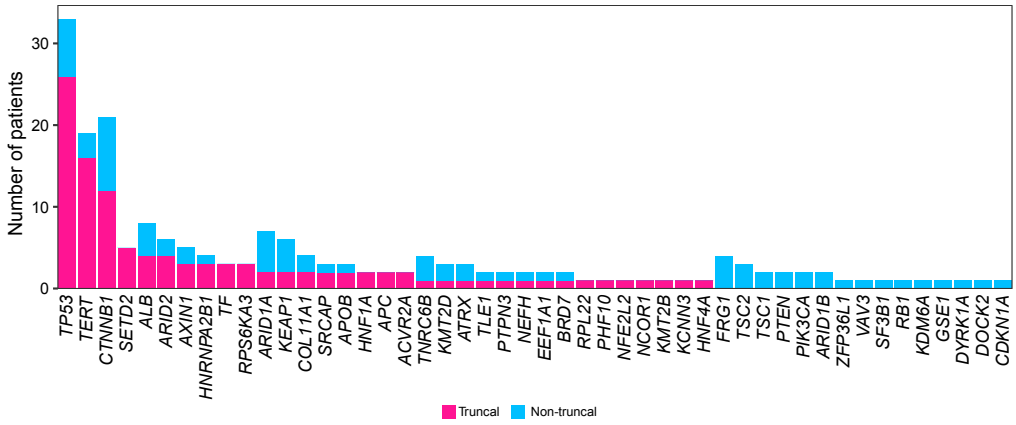


Figure 4.6: Clonality status of drivers in the PLANET cohort. The barplot indicates the number of patients with truncal and non-truncal mutations in each driver gene. If the driver is mutated across all sectors of a tumor, this is denoted as a truncal mutation. Classical drivers such as TP53, TERT and CTNNB1 are mostly truncal but non-truncal mutations are also observed. Some genes only have non-truncal mutations such as FRG1 but these genes are not very frequent in the PLANET cohort

approach. In summary, even though classical driver mutations arise early in the history of tumorigenesis, there are significant amount of subclonal driver mutations across many HCCs, empowering clonal expansion and evolution.

4.3.5 Timing of mutational processes

Signature deconvolution in truncal and non-tuncal mutations can illuminate the evolution of mutational process in HCC. In addition to the timing of driver alterations, the mutational signatures across the history of the tumorigenesis were compared. Signatures related to external stimulus such as aristolochic acid (SBS22), smoking (SBS4) and aflatoxin B1 (SBS24) are much higher in the early (truncal) part of the evolution (Figure 4.8), implying that

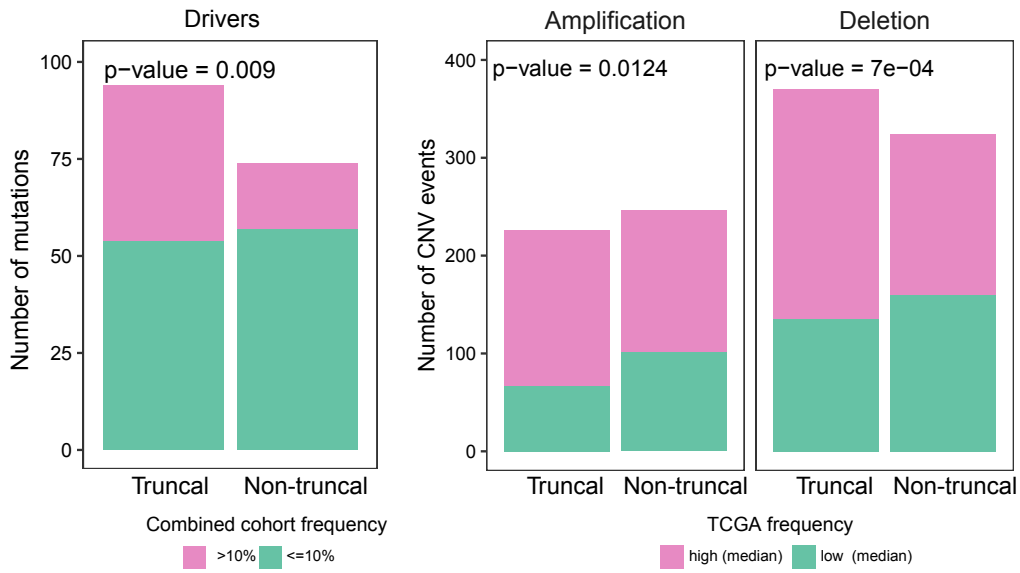


Figure 4.7: Mutation freqeuncy and timing comparison. Number of truncal and non-truncal driver mutations (left), amplifications (middle) and deletions (right) partitioned by their occurrence level in TCGA cohort. Driver gene mutations, amplifications and deletions with a low population frequency tend to be non-truncal alterations.

these mutational processes have accompanied tumor initiation. In addition, increase in proportions of age related SBS1 and SBS5 signatures were observed (Figure 4.8).

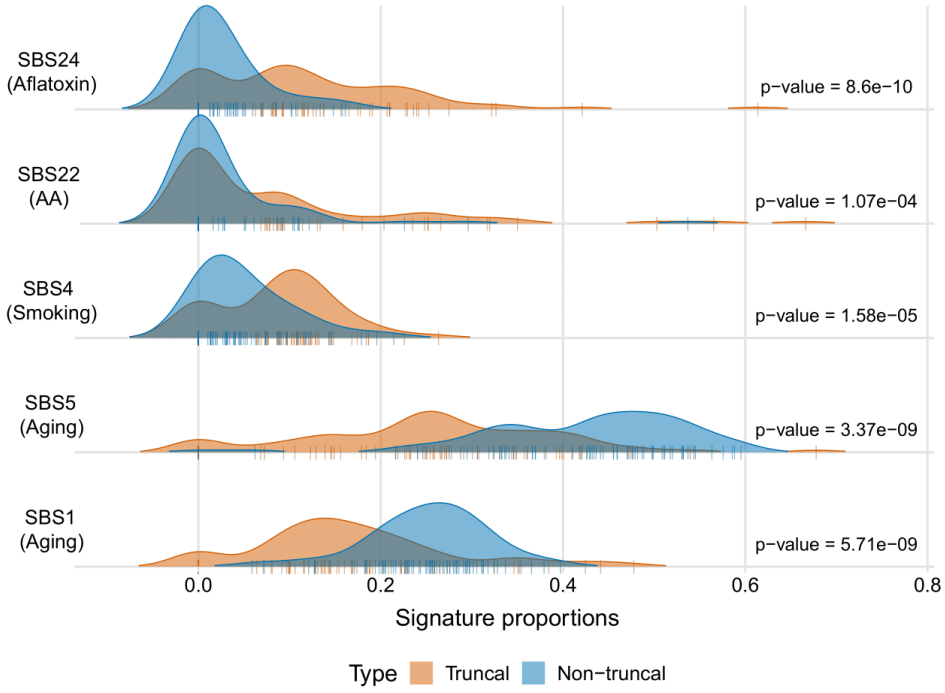


Figure 4.8: Truncal and non-truncal signature proportions. Signature proportions were calculated seperately on truncal (shared by all sectors) and non-truncal mutations and proportion distributions are compared using paired Wilcoxon signed-rank test. Signatures which show significant difference in proportions are displayed. While signatures related to external exporsures such as smoking and aristolochic acid (AA) tend to enrich in early (truncal) mutations, this pattern is reverse for age related signatures.

4.3.6 Large scale copy number events arise early

Using genome wide sequencing depth as well as the frequency of the germline variants, we inferred the copy number alterations (CNA) and the purity of the

samples (Favero et al., 2015). Among chromosomal-level copy number events, amplifications at 8q (66%), 1q (68%) and 7p (38%) together with deletions at 8p (77%), 17p (60%), 16q (54%) and 4q (46%) were the most frequent arm-level events (Figure 4.9). Arm-level amplifications and deletions are not only very similar across multiple sectors of the same patient, but also concordant with the TCGA cohort when comparing with the TCGA Asian cohort (Figure 4.9). Majority of frequent arm level amplifications and deletions are truncal events (Figure 4.10). These suggest that large chromosomal events tend to be early events in the history of tumorigenesis (Gao et al., 2019).

In addition to large scale copy number changes, driver CNAs were identified using the TCGA Asian cohort was conducted, (Chapter 3, GISTIC analysis). Several focal CNAs containing cancer related genes such as *TERT* and *MYC* were recurrently amplified, while tumor suppressor genes such as *RB1* and *CASP3* were frequently deleted (Figure 4.11). Despite many shared large scale chromosomal CNVs, focal CNVs are often subclonal in many patients, even though high frequency amplification and deletions across patients also tend to be truncal events within individual patient (p = 0.0124 and p = 7e-04, Figure 4.7 middle and right). Taken together, these observations suggest that while large-scale copy number events are often shared across sectors, there are active gain and loss of focal CNAs driving further diversification of each tumor.

4.3.7 Variable amount of DNA ITH in the PLANET cohort

Multiple regions from a tumor also allows us to identify the intra-tumor heterogeneity (ITH) using all mutations. In this study, 2 to 11 sectors were

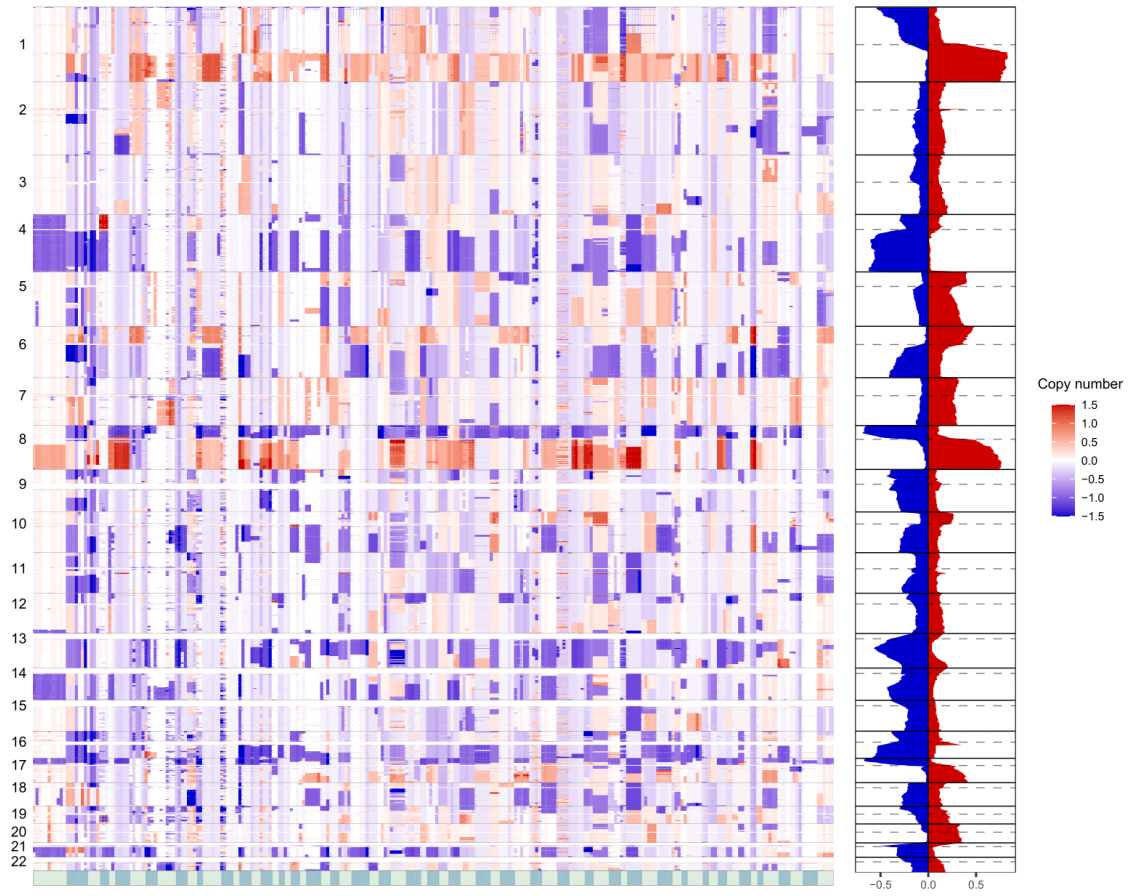


Figure 4.9: Genome-wide copy number landscape of the cohort across samples. Samples fro mthe same patient were indicated using a bar at the bottom og the heatmap with alternating colors. Large copy number alterations tend to be shared across all sectors of patients. The plot at the right shows the amplification/deletion proportions of corresponding chromosome arms in the TCGA Asian cohort. Recurrent events in the PLANET cohort is compatible with public cohort.

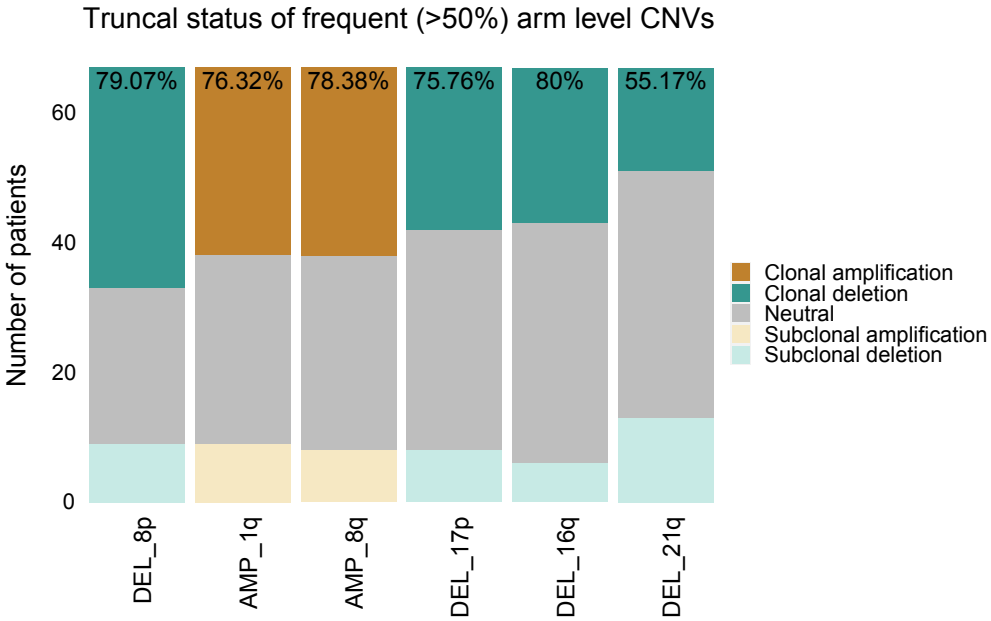


Figure 4.10: Proportions of truncal and non-truncal arm events. Frequency of truncal arm level event proportions are shown for frequent amplifications and deletions. Majority of frequent arm level events are clonal.

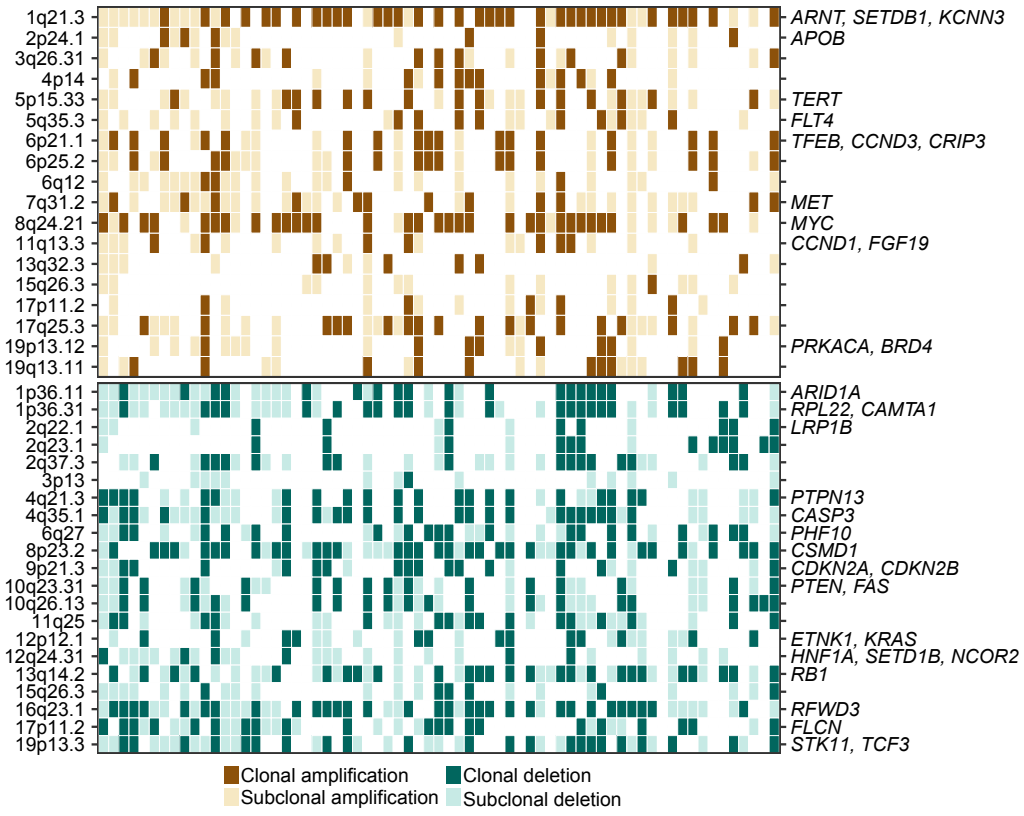


Figure 4.11: Clonal status of amplification (top) and deletions (bottom) in cytobands with significant CNV events. Lighter colors represent subclonal amplifications and deletions. Genes located at these cytobands are labeled if the event is amplification and the role of gene is oncogene or if the event is deletion and the role of gene is tumor suppressor based on the COSMIC database. Many cytoband level copy number events are non-truncal.

samples from each tumor with a median number of sectors of 4. Using mutation presence and absence status across the genome, the mean value across all pairwise distances tumor sectors were calculated to measure the overall DNA ITH for each patient (See methods). DNA ITH varied greatly between 14% and 96% across patients with a mean DNA ITH of 42% and a standard deviation of 16%. When I plot the tree of phylogenetic relationships across tumor sectors, low DNA ITH patients showed a long trunk which indicates many shared mutations across sectors while trees of high DNA ITH patients showed shorts trunk and long branches (Figure 4.12).

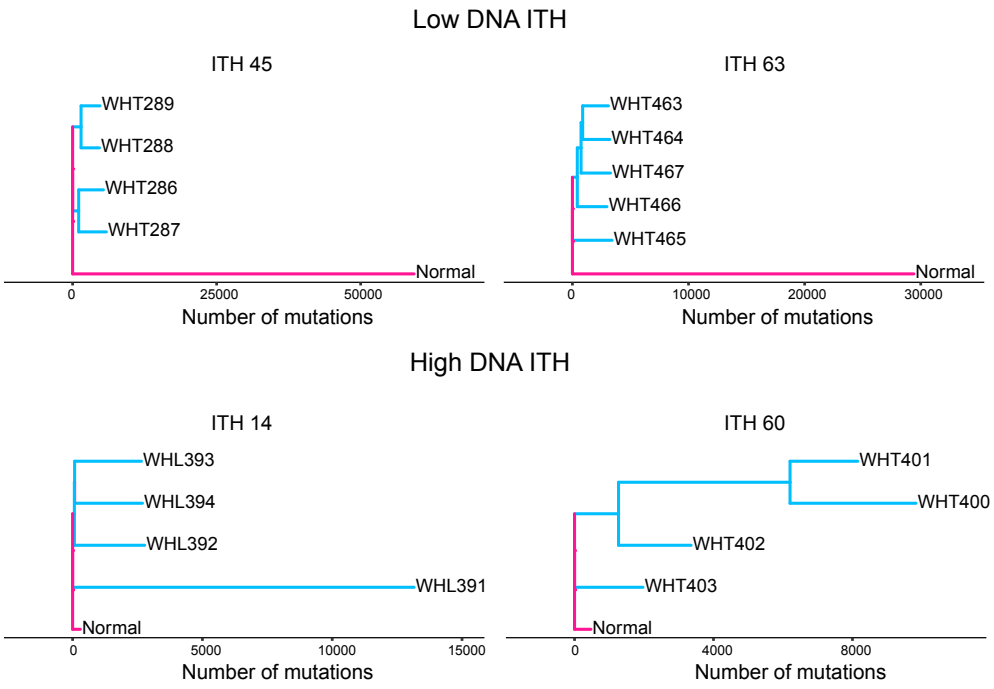


Figure 4.12: Example DNA trees for low and high ITH patients. Trees of two tumors with lowest DNA ITH and two tumors with the highest DNA ITH are shown.

4.3.8 Co-existing RNA subtypes in the PLANET cohort

Using the non-negative matrix factorization, I clustered the RNA subtypes across 198 sectors with the RNAseq data. Optimum number of clusters were 3 or 4 based on cophenetic correlation coefficient as well as the silhouette value (Figure 4.13 A-B). As four RNA subtypes for the Asians were identified in chapter 3, four was used as the number of subtypes. In order to understand the concordance across subtypes between the PLANET and the TCGA Asian cohort, gene contributions (weights) to each subtype were correlated using NMF outputs. Each of the PLANET subtypes showed a significant correlation to the subtypes from the TCGA Asian subtypes (Figure 4.13 C). While spearman correlation is usually quite high between corresponding subtypes, P2 subtype showed the lowest correlation with the Asian P2 subtype (rho=0.41) potentially due to small sample size and lower number of P2 subtype in the PLANET cohort (n=4, Figure 4.13 C). Subtypes were then named based on the corresponding Asian subtypes.

With multiple sectors from each tumor, degree of transcriptomic heterogeneity can also be measured using two different approaches: 1) we can ask whether a tumor has sectors belonging to different RNA subtypes or 2) we can interrogate the transcriptomic distance between sectors of the tumor using all expression values. Interestingly, 18 patients have tumors sectors belonging to multiple RNA subtypes (Figure 4.14).

In addition to mixed subtypes, overall level of RNA ITH was also calculated using the expression levels of all protein coding genes. RNA ITH was also varied

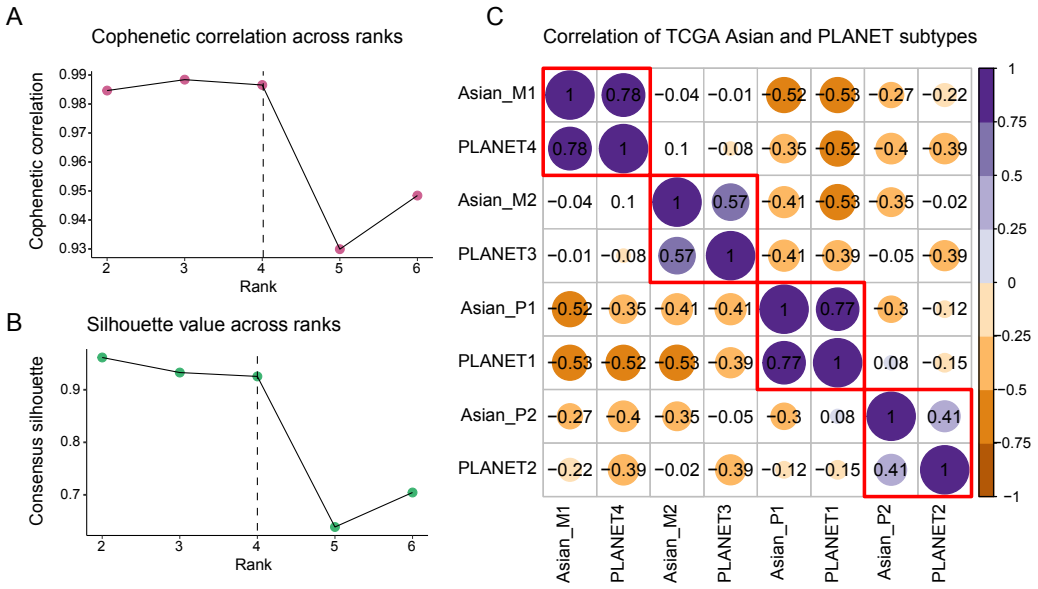


Figure 4.13: RNA subtypes in the PLANET cohort. (A) Cophenetic correlation coefficient across ranks 2:6 (B) Silhouette value across ranks 2:6. Bases on both metrics, 4 subtypes can be selected for the PLANET cohort. (C) Correlation of PLANET subtypes to TCGA Asian cohort subtypes. Spearman correlation was calculated using NMF weights of shared genes among top 3000 most variable genes that were used in NMF analysis for both cohorts. Only significant correlations are colored. All pairwise correlation coefficients are then clustered. All subtypes in TCGA Asian cohort clusters with a PLANET cohort subtype. Subtypes in the PLANET cohort were named based on similarity to TCGA Asian cohort subtypes.

across patients. Interestingly, while patients with mixed subtypes showed relatively higher level of RNA ITH, this difference was close to statistical significance (p = 0.06). This implies that overall transcriptomic heterogeneity does not always reflect the functional heterogeneity.

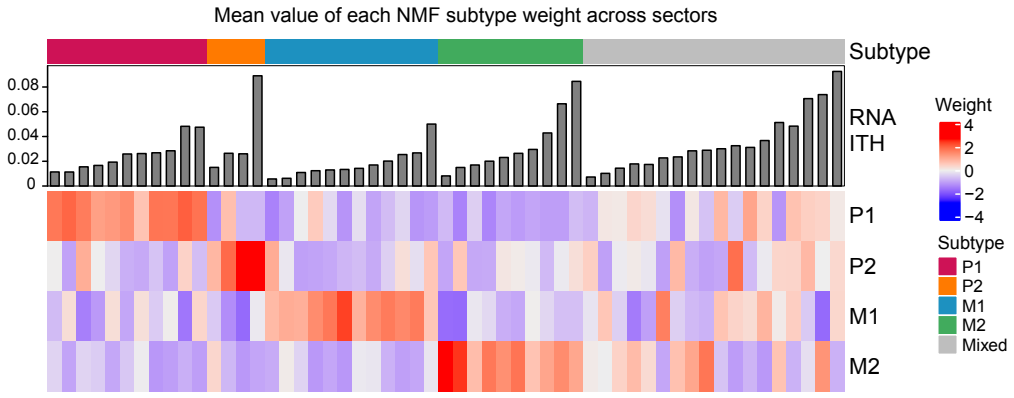


Figure 4.14: Mixed RNA subtypes in the PLANET cohort. Heatmap shows the mean NMF weights for each subtype across the sectors of each patient. Subtype is determined based on the highest NMF weight and assigned subtypes are shown at the top annotation. 18 patients have sectors from multiple subtypes (gray). RNA ITH is shown as barplot annotation at the top of the heatmap.

4.3.9 ITH and clinical outcome

In addition to levels of ITH, one important question is how ITH observed at the DNA, RNA as well as the immune level correlate with information from other layers (e.g. driver status) and whether ITH can stratify patient in terms of their prognosis and survival. In order to explore correlation among multiple features, I collected basic clinical features (n=4 e.g. stage), molecular features (n=8, e.g. RNA subgroups and immune subtypes) as well as ITH features (DNA, RNA and immune ITH). When plotting the correlation among multiple

layers, correlation exists across features from multiple layers and ITH features correlated with multiple features within and between categories (Figure 4.15).

Using relapse free survival for this prospective cohort, a multivariate cox model was generated to test how features across multiple layers can be combined in predicting patient survival (Figure 4.16). Despite they are not significant in the univariate analysis (Figure 4.15), RNA ($p < 0.001$) and DNA ITH ($p < 0.001$) features contribute strongly to the univariate cox model, suggesting the importance of collecting ITH features for future patient management and stratification (Figure 4.16). When ranking features based an their contribution the the Cox model, clinical features stage and age rank first and DNA and RNA ITH values have intermediate importance (Figure 4.17). Thus, intra-tumor heterogeneity can contribute significantly to patient survival and prognosis.

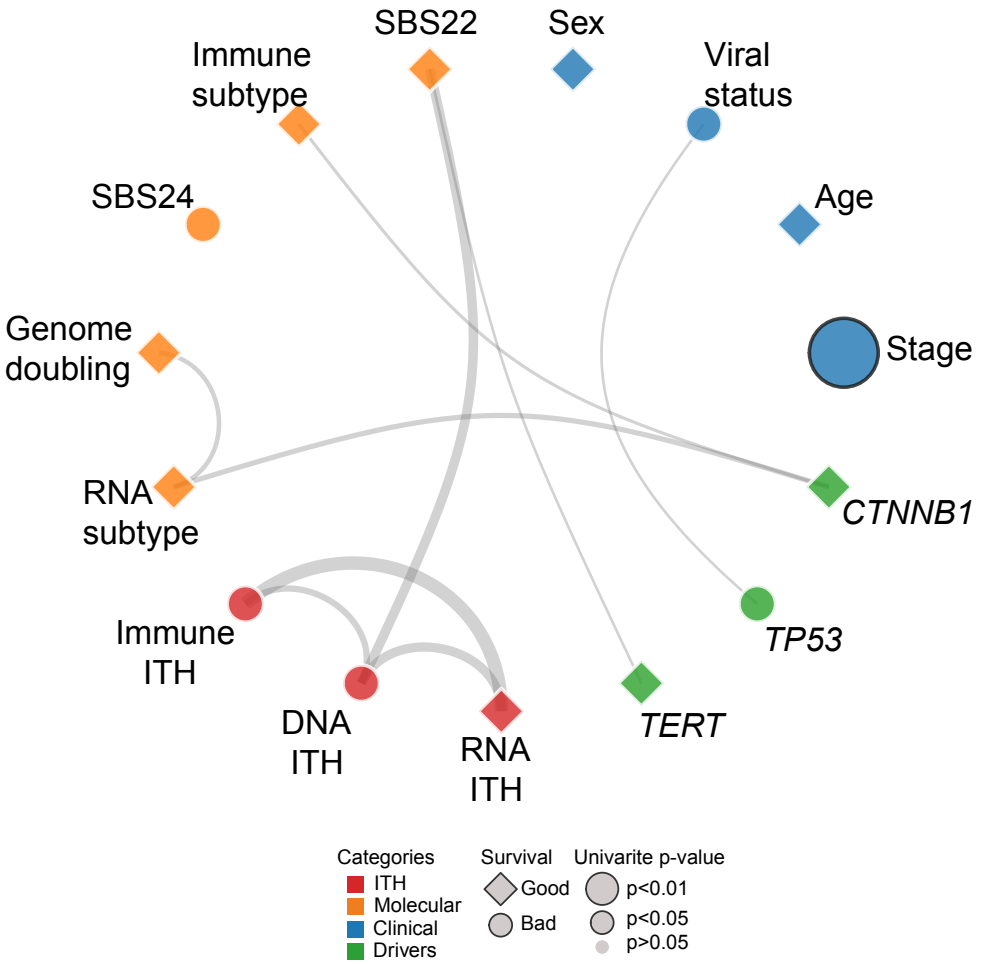


Figure 4.15: Correlation network and univariate survival analysis of selected features from clinical, molecular and ITH categories. Edges of the network indicates correlation between features at the connected nodes. The thickness of edges are re-scaled p-values after -log10(p) transformation. The diamond shape represents a hazard ratio (HR) less than 1 (later recurrence) and circles represent a HR greater than 1 (earlier recurrence). For features with multiple levels such as Stage, HR of the most significant level is used. The black border around the triangle and the bigger size means that feature has a Cox model log-rank score test p-value<0.05 in the univariate analysis.

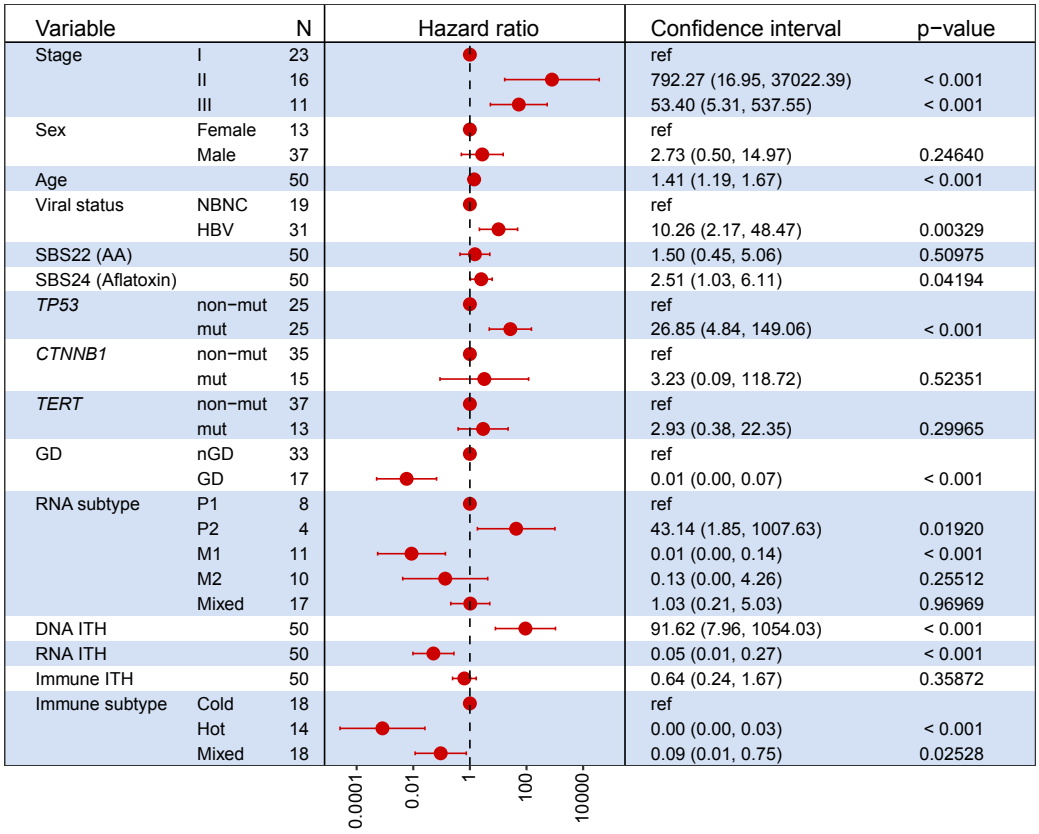


Figure 4.16: Forest plot of multivariate Cox model for the PLANET cohort. Levels for each features and numbers are shown. Hazard ratio as well as confidence intervals of hazard ratio is also plotted. DNA and RNA ITH features are significant in this multivariate survival model.

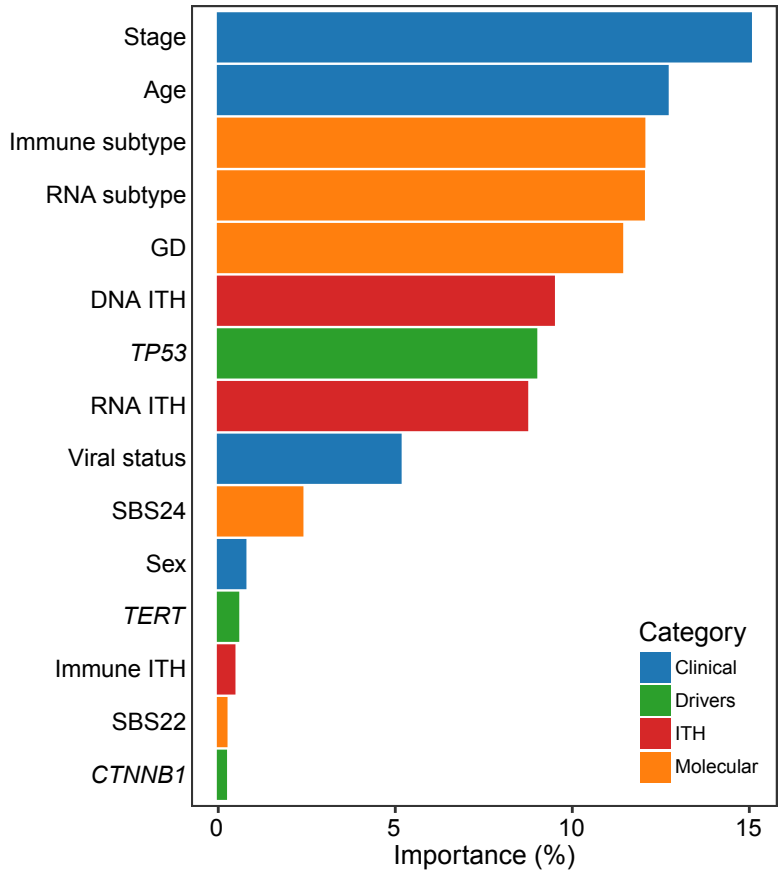


Figure 4.17: Importance ranking of variables. Bars show the percentage of chi-squared value of variable from likelihood ratio test in the sum of all chi-squared values. While clinical features such as stage and age rank first, DNA and RNA ITH features have intermediate importance for survival prediction.

4.4 Discussions

4.4.1 Selection criteria for analysis dataset

While variable levels of intratumor heterogeneity (ITH) have been revealed by multi-region studies for different cancer types such as lung (Jamal-Hanjani et al., 2017) and renal cell carcinoma (Gerlinger et al., 2014), ITH in HCC is limited to studies with small sample size (Xue et al., 2016; Zhai et al., 2017). PLANET is a unique dataset which multiple sectors (n=2-11) from 67 patients were sequenced in the Genome Institute of Singapore. With 318 whole genome sequencing and 253 RNAseq data, this dataset is the unique multi-sectoring HCC dataset which enables intratumor heterogeneity analysis at DNA and RNA level.

4.4.2 Rationales for bioinformatics tool selection

Mutational signature analysis: As discussed in Section 3.3.2, mutational signature analysis can be done either by discovering signatures de-novo (e.g. BayesNMF by Kasar et al. (2015)) 2) and by desconstructing to known signatures (e.g. deconstrucsigs by Rosenthal et al. (2016)). As PLANET cohort genome was sequenced using WGS technology, we have the full set of mutation info which indicates that de-novo mutation identification is employable. In this chapter, de-novo signature analysis was implemented using BayesNMF tool (Kasar et al., 2015). Interestingly, 7 out of 9 de-novo signatures identified in this cohort were largely similar to signatures identified by previous studies which I also used in Chapter 3. SBS25 (chemotherapy) and SBS40 (unknown) are two novel signatures identified in this cohort.

Identification of RNA subtypes: Using non-negative matrix factorization has benefits as discussed in Section 3.3.2 such as assigned gene weights based on their contribution to the subtype. As, this provides opportunity to compare subtypes in this cohort to the subtypes in previous analysis (in Chapter 3), NMF algorithm was used to identify RNA subtypes.

4.4.3 Comparison of findings with the literature

Previous ITH studies identified variable levels of intratumor heterogeneity (ITH) in HCC similar to our finding in the PLANET cohort. Although their sequencing is targeted and not as comprehensive as our analysis (WGS), Friemel et al. (2015) observed mutational heterogeneity in *TP53* and *CTNNB1* genes in 22% in a cohort of 23 HCC patients. In addition, Xue et al. (2016) also reported very variable landscape of ITH in HCC where the percentage of mutations shared across all tumors sectors varied greatly (8%-97%) in a cohort of 10 patients. Similarly, Zhai et al. (2017) reported similar variable ITH using a small subset of patients (n=9) from the PLANET cohort. Although findings on ITH in this study converges with the literature findings, PLANET cohort has the biggest number of samples and provides a reliable landscape of ITH in HCC.

The study of the natural history of HCC through a unique prospective HCC cohort PLANET and provided several novel insights. Firstly, many subclonal driver mutations across the cohort was observed. Subclonal driver can drive continuous evolution of the tumor which might result in treatment failure. Strikingly, 60% chromatin remodeling genes such as *ARID1A*, *ARID2* and

KDM6A are subclonal mutations, while only 38% of mutations in other drivers are subclonal (p = 0.04, Figure 4.18, Figure 4.6). This confirms the findings in Chapter 2 that chromatin remodeling pathway is late in the history of tumorigenesis (Figure 2.13 E). In addition to point mutations and indels in drivers, frequent focal copy number alterations such as amplifications in *FGF19* were reported for HCC (Ahn et al., 2014; Shibata & Aburatani, 2014). Recently, a phase 1 clinical trial revealed the efficacy of fisogatinib drug on 17% of HCCs with *FGF19* aberrations but no response was observed for *FGF19* wild-type tumors (Kim et al., 2019). With the multi-region data of the PLANET cohort, many subclonal focal drivers are revealed. Interestingly, while 30% of PLANET cohort had *FGF19* amplifications, only one third of the patients had truncal events in this gene (Figure 4.11), which might explain why not all *FGF19* positive tumors responded to fisoganatib. Thus, with this comprehensive atlas, we were able to document the level of ITH across a wide range of patients and these information can provide important basis designing better strategies stratifying patients for treatments.

We observed co-existence of multiple molecular subtypes within a tumor (Figure 4.14). Multiple RNA subtypes observed in patients might be the reason for treatment failure in HCC. Understanding the drivers of these mixed RNA subtypes will help to understand the origin of mixed subtype and potentially overcome the treatment failure. As mutation in chromatin remodeling driver genes might cause phenotypic plasticity (Flavahan et al., 2017; Virk et al., 2020), we did observe a slight enrichment of these mutations in mixed subtype patients (28% in the mixed subtype patients vs 13% in the

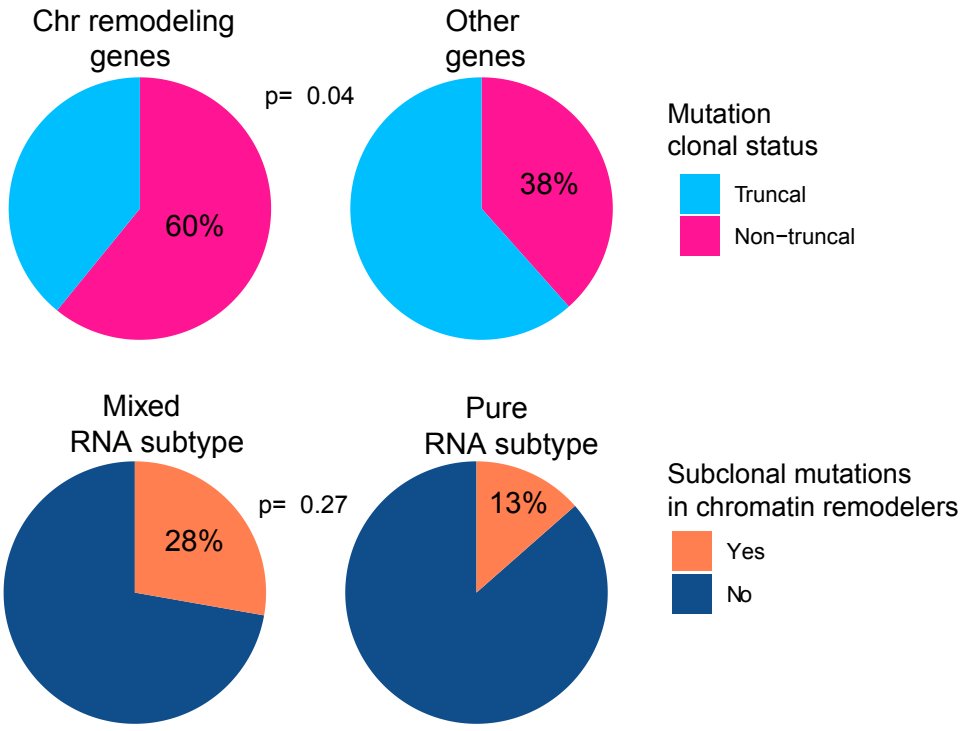


Figure 4.18: Chromatin remodeling genes and mixed subtypes. Top charts show non-truncal mutation percentages in chromatin remodeling genes. Bottom charts show the percentage of patients with non-truncal mutations in chromatin remodeling genes in mixed and pure rna subtypes. Non-truncal mutations enriche in chromatin remodeling genes. Patients with mixed RNA subtypes have higher proportion of chromatin remodeling genes with a subclonal mutation but this is not statistically significant.

background sequences, p = 0.27, Figure 4.18). Different from our finding, Hayashi et al. (2020) identified a link between clonal mutations in chromatin remodeling genes of pancreatic cancer and existence of a subclonal molecular subtype with squamous features. This might imply that epigenetic plasticity is gained early in pancreatic cancer but late in HCC. Understanding the inter-relationship between genomic changes and phenotypic evolution will be a key direction for the field.

4.4.4 Evolution of mutational signatures

By dissecting the mutational process in the trunk and branches, we identified a noticable decrease in the contribution of aristolochic acid (AA, SBS22), smoking (SBS4) and aflatoxin B1 (SBS24) signatures, all of which are linked to environmental exposures (Figure 4.8). In chapter 2, we also observed similar decrease in AA and smoking signatures in both Asian and European cohorts. This might be due to attenuation of exposure to environment related carcinogen (e.g. AA, smoking or aflatoxin). The increase in the proportions of age-related signatures and decrease in aflatoxin B1 signature were also found in a previous European study (Letouzé et al., 2017).

4.4.5 Predictive importance of ITH

Integrative survival analysis revealed that both DNA and RNA ITH are contributiong to relapse-free survival significantly (Figure 4.16). Interestingly, while high DNA ITH ($p < 0.0001$) is associated with earlier relapse, the reverse is true for RNA ITH ($p < 0.0001$). While higher DNA ITH might contribute to

the emergence of fitter subclone which can contribute to patient relapse, higher RNA ITH might have protective effects. It is worth noting that, despite slight correlation, RNA ITH is not significantly correlated to mixed RNA subtypes (Figure 4.14). This implies that heterogeneity in certain pathways might be enough to convert to a new subtype.

Publications of the author

Nguyen, P., Ma, S., Phua, C., **Kaya, N.**, Lai, H., & Lim, C. et al. (2021). Intratumoural immune heterogeneity as a hallmark of tumour evolution and progression in hepatocellular carcinoma. Nature Communications, 12(1). doi: 10.1038/s41467-020-20171-7

Poster presentations of the author

Kaya, N., Zhai, W. (2018). An Evolutionary and Genomic Approach to Understanding Tumor Evolution in Hepatocellular Carcinoma. Evolutionary Biology and Ecology of Cancer Summer School, Wellcome Genome Campus, Cambridge, UK.

References

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., & Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nature Methods*, *7*, 248. https://doi.org/10.1038/nmeth0410-248

Ahn, S.-M., Jang, S. J., Shim, J. H., Kim, D., Hong, S.-M., Sung, C. O., Baek, D., Haq, F., Ansari, A. A., Lee, S. Y., Chun, S.-M., Choi, S., Choi, H.-J., Kim, J., Kim, S., Hwang, S., Lee, Y.-J., Lee, J., Jung, W., ... Kong, G. (2014). Genomic portrait of resectable hepatocellular

carcinomas: Implications of RB1 and FGF19 aberrations for patient stratification. *Hepatology*, *60*(6), 1972–1982. https://doi.org/10.1002/hep.27198

Alkallas, R., Lajoie, M., Moldoveanu, D., Hoang, K. V., Lefrançois, P., Lingrand, M., Ahanfeshar-Adams, M., Watters, K., Spatz, A., Zippin, J. H., Najafabadi, H. S., & Watson, I. R. (2020). Multi-omic analysis reveals significantly mutated genes and DDX3X as a sex-specific tumor suppressor in cutaneous melanoma. *Nature Cancer*, *1*(6), 635–652. https://doi.org/10.1038/s43018-020-0077-8

Ally, A., Balasundaram, M., Carlsen, R., Chuah, E., Clarke, A., Dhalla, N., Holt, R. A., Jones, S. J. M., Lee, D., Ma, Y., Marra, M. A., Mayo, M., Moore, R. A., Mungall, A. J., Schein, J. E., Sipahimalani, P., Tam, A., Thiessen, N., Cheung, D., … Laird, P. W. (2017). Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell*, *169*(7), 1327–1341.e23. https://doi.org/10.1016/j.cell.2017.05.046

Anagnostou, V., Niknafs, N., Marrone, K., Bruhm, D. C., White, J. R., Naidoo, J., Hummelink, K., Monkhorst, K., Lalezari, F., Lanis, M., Rosner, S., Reuss, J. E., Smith, K. N., Adleff, V., Rodgers, K., Belcaid, Z., Rhymee, L., Levy, B., Feliciano, J., … Velculescu, V. E. (2020). Multimodal genomic features predict outcome of immune checkpoint blockade in non-small-cell lung cancer. *Nature Cancer*, *1*(1), 99–111. https://doi.org/10.1038/s43018-019-0008-8

Anani, W., & Shurin, M. R. (2017). Targeting Myeloid-Derived Suppressor Cells in Cancer. *Advances in Experimental Medicine and Biology*, *1036*, 105–128. https://doi.org/10.1007/978-3-319-67577-0_8

Andor, N., Graham, T. A., Jansen, M., Xia, L. C., Aktipis, C. A., Petritsch, C., Ji, H. P., & Maley, C. C. (2016). Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nature Medicine*, *22*(1), 105–+. https://doi.org/10.1038/nm.3984

Bien, S. A., & Peters, U. (2019). Moving from one to many: Insights from the growing list of pleiotropic cancer risk genes. *British Journal of Cancer*, *120*(12), 1087–1089. https://doi.org/10.1038/s41416-019-0475-9

Birkbak, N. J., Eklund, A. C., Li, Q., McClelland, S. E., Endesfelder, D., Tan, P., Tan, I. B., Richardson, A. L., Szallasi, Z., & Swanton, C. (2011). Paradoxical Relationship between Chromosomal Instability and Survival Outcome in Cancer. *Cancer Research*, *71*(10), 3447–3452. https://doi.org/10.1158/0008-5472.can-10-3667

Boveri, T. (2008). Concerning the Origin of Malignant Tumours by Theodor Boveri. Translated and annotated by Henry Harris. *Journal of Cell Science*, *121*(Supplement 1), 1. https://doi.org/10.1242/jcs.025742

Boyault, S., Rickman, D. S., Reyniès, A. de, Balabaud, C., Rebouissou, S., Jeannot, E., Hérault, A., Saric, J., Belghiti, J., Franco, D., Bioulac-Sage, P., Laurent-Puig, P., & Zucman-Rossi, J. (2007). Transcriptome classification of HCC is related to gene alterations and to new therapeutic targets. *Hepatology*, *45*(1), 42–52. https://doi.org/10.1002/hep.21467

Bozic, I., & Wu, C. J. (2020). Delineating the evolutionary dynamics of cancer from theory to reality. *Nature Cancer*, *1*(6), 580–588. https://doi.org/10.1038/s43018-020-0079-6

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *68*(6), 394–424. https://doi.org/10.3322/caac.21492

Chaisaingmongkol, J., Budhu, A., Dang, H., Rabibhadana, S., Pupacdi, B., Kwon, S. M., Forgues, M., Pomyen, Y., Bhudhisawasdi, V., Lertprasertsuke, N., Chotirosniramit, A., Pairojkul, C., Auewarakul, C. U., Sricharunrat, T., Phornphutkul, K., Sangrajrang, S., Cam, M., He, P., Hewitt, S. M., ... Wang, X. W. (2017). Common Molecular Subtypes Among Asian Hepatocellular Carcinoma and Cholangiocarcinoma. *Cancer Cell*. https://doi.org/10.1016/j.ccell.2017.05.009

Chaudhary, K., Poirion, O. B., Lu, L., Huang, S., Ching, T., & Garmire, L. X. (2018). Multi-modal meta-analysis of 1494 hepatocellular carcinoma samples reveals significant impact of consensus driver genes on phenotypes. *Clinical Cancer Research*, clincanres.0088.2018. https://doi.org/10.1158/1078-0432.CCR-18-0088

Chen, C., & Wang, G. (2015). Mechanisms of hepatocellular carcinoma and challenges and opportunities for molecular targeted therapy. *World Journal of Hepatology*, *7*(15), 1964–1970. https://doi.org/10.4254/wjh.

v7.i15.1964

Chen, J., Yang, H., Teo, A. S. M., Amer, L. B., Sherbaf, F. G., Tan, C. Q., Alvarez, J. J. S., Lu, B., Lim, J. Q., Takano, A., Nahar, R., Lee, Y. Y., Phua, C. Z. J., Chua, K. P., Suteja, L., Chen, P. J., Chang, M. M., Koh, T. P. T., Ong, B.-H., ... Zhai, W. (2020). Genomic landscape of lung adenocarcinoma in East Asians. *Nature Genetics*. https://doi.org/10.1038/s41588-019-0569-6

Chiang, D. Y., Villanueva, A., Hoshida, Y., Peix, J., Newell, P., Minguez, B., LeBlanc, A. C., Donovan, D. J., Thung, S. N., Sole, M., Tovar, V., Alsinet, C., Ramos, A. H., Barretina, J., Roayaie, S., Schwartz, M., Waxman, S., Bruix, J., Mazzaferro, V., ... Llovet, J. M. (2008). Focal Gains of Vascular Endothelial Growth Factor A and Molecular Classification of Hepatocellular Carcinoma. *Cancer Research*, *68*(16), 6779–6788. https://doi.org/10.1158/0008-5472.CAN-08-0742

Chin, L., Andersen, J. N., & Futreal, P. A. (2011). Cancer genomics: From discovery science to personalized medicine. *Nat Med*, *17*(3), 297–303. https://doi.org/10.1038/nm.2323

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S., & Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, *31*, 213. https://doi.org/10.1038/nbt.2514

Cleary, S. P., Jeck, W. R., Zhao, X., Chen, K., Selitsky, S. R., Savich, G. L., Tan, T.-X., Wu, M. C., Getz, G., Lawrence, M. S., Parker, J. S., Li, J., Powers, S., Kim, H., Fischer, S., Guindi, M., Ghanekar, A., & Chiang, D. Y. (2013). Identification of driver genes in hepatocellular carcinoma by exome sequencing. *Hepatology*, *58*(5), 1693–1702. https://doi.org/10.1002/hep.26540

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M. R., Jassal, B., Jupe, S., Matthews, L., May, B., Palatnik, S., Rothfels, K., Shamovsky, V., Song, H., Williams, M., ... D'Eustachio, P. (2014). The Reactome pathway knowledgebase. *Nucleic Acids Research*, *42*(D1), D472–D477. https://doi.org/10.1093/nar/gkt1102

Danaher, P., Warren, S., Dennis, L., D'Amico, L., White, A., Disis, M. L., Geller, M. A., Odunsi, K., Beechem, J., & Fling, S. P. (2017). Gene expression markers of Tumor Infiltrating Leukocytes. *Journal for ImmunoTherapy of Cancer*, *5*(1), 18. https://doi.org/10.1186/s40425-017-0215-8

Davoli, T., Uno, H., Wooten, E. C., & Elledge, S. J. (2017). Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science*, *355*(6322), eaaf8399. https://doi.org/10.1126/science.aaf8399

Davoli, T., Xu, A. W., Mengwasser, K. E., Sack, L. M., Yoon, J. C., Park, P. J., & Elledge, S. J. (2013). Cumulative Haploinsufficiency

and Triplosensitivity Drive Aneuploidy Patterns and Shape the Cancer Genome. *Cell*, *155*(4), 948–962. https://doi.org/10.1016/j.cell.2013.10.011

De Cicco, P., Ercolano, G., & Ianaro, A. (2020). The New Era of Cancer Immunotherapy: Targeting Myeloid-Derived Suppressor Cells to Overcome Immune Evasion. *Frontiers in Immunology*, *11*. https://doi.org/10.3389/fimmu.2020.01680

Dees, N. D., Zhang Q Fau - Kandoth, C., Kandoth C Fau - Wendl, M. C., Wendl Mc Fau - Schierding, W., Schierding W Fau - Koboldt, D. C., Koboldt Dc Fau - Mooney, T. B., Mooney Tb Fau - Callaway, M. B., Callaway Mb Fau - Dooling, D., Dooling D Fau - Mardis, E. R., Mardis Er Fau - Wilson, R. K., Wilson Rk Fau - Ding, L., & Ding, L. (2012). *MuSiC: Identifying mutational significance in cancer genomes. 1549-5469 (Electronic).*

Deng, J., Chen, H., Zhou, D., Zhang, J., Chen, Y., Liu, Q., Ai, D., Zhu, H., Chu, L., Ren, W., Zhang, X., Xia, Y., Sun, M., Zhang, H., Li, J., Peng, X., Li, L., Han, L., Lin, H., ... Zhao, K. (2017). Comparative genomic analysis of esophageal squamous cell carcinoma between Asian and Caucasian patient populations. *Nature Communications*, *8*(1), 1533. https://doi.org/10.1038/s41467-017-01730-x

Ding, L., Bailey, M. H., Porta-Pardo, E., Thorsson, V., Colaprico, A., Bertrand, D., Gibbs, D. L., Weerasinghe, A., Huang, K., Tokheim, C., Cortés-Ciriano, I., Jayasinghe, R., Chen, F., Yu, L., Sun, S., Olsen,

C., Kim, J., Taylor, A. M., Cherniack, A. D., ... Mariamidze, A. (2018). Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell*, *173*(2), 305–320.e10. https://doi.org/10.1016/j.cell.2018.03.033

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2012). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dow, M., Pyke, R. M., Tsui, B. Y., Alexandrov, L. B., Nakagawa, H., Taniguchi, K., Seki, E., Harismendy, O., Shalapour, S., Karin, M., Carter, H., & Font-Burgada, J. (2018). Integrative genomic analysis of mouse and human hepatocellular carcinoma. *Proceedings of the National Academy of Sciences*, *115*(42), E9879. https://doi.org/10.1073/pnas.1811029115

Favero, F., Joshi, T., Marquard, A. M., Birkbak, N. J., Krzystanek, M., Li, Q., Szallasi, Z., & Eklund, A. C. (2015). Sequenza: Allele-specific copy number and mutation profiles from tumor sequencing data. *Annals of Oncology*, *26*(1), 64–70. https://doi.org/10.1093/annonc/mdu479

Flavahan, W. A., Gaskell, E., & Bernstein, B. E. (2017). Epigenetic plasticity and the hallmarks of cancer. *Science*, *357*(6348), eaal2380. https://doi.org/10.1126/science.aal2380

Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C. G., Ward, S., Dawson, E., Ponting, L., Stefancsik, R., Harsha, B., Kok, C. Y., Jia, M., Jubb, H., Sondka, Z., Thompson, S.,

De, T., & Campbell, P. J. (2017). COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Research*, *45*(D1), D777–D783. https://doi.org/10.1093/nar/gkw1121

Forner, A., Reig, M., & Bruix, J. (2018). Hepatocellular carcinoma. *The Lancet*, *391*(10127), 1301–1314. https://doi.org/10.1016/S0140-6736(18)30010-2

Friemel, J., Rechsteiner, M., Frick, L., Böhm, F., Struckmann, K., Egger, M., Moch, H., Heikenwalder, M., & Weber, A. (2015). Intratumor Heterogeneity in Hepatocellular Carcinoma. *Clinical Cancer Research*, *21*(8), 1951. https://doi.org/10.1158/1078-0432.CCR-14-0122

Fujimoto, A., Furuta, M., Totoki, Y., Tsunoda, T., Kato, M., Shiraishi, Y., Tanaka, H., Taniguchi, H., Kawakami, Y., Ueno, M., Gotoh, K., Ariizumi, S., Wardell, C. P., Hayami, S., Nakamura, T., Aikata, H., Arihiro, K., Boroevich, K. A., Abe, T., ... Nakagawa, H. (2016). Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat Genet*, *48*(5), 500–509. https://doi.org/10.1038/ng.3547

Galarreta, M. R. de, Bresnahan, E., Molina-Sánchez, P., Lindblad, K. E., Maier, B., Sia, D., Puigvehi, M., Miguela, V., Casanova-Acebes, M., Dhainaut, M., Villacorta-Martin, C., Singhi, A. D., Moghe, A., Felden, J. von, Grinspan, L. T., Wang, S., Kamphorst, A. O., Monga, S. P., Brown, B. D., ... Lujambio, A. (2019). β-Catenin Activation Promotes Immune Escape and Resistance to Anti–PD-1 Therapy in Hepatocellular

Carcinoma. *Cancer Discovery*, *9*(8), 1124–1141. https://doi.org/10.1158/2159-8290.CD-19-0074

Gao, Q., Zhu, H., Dong, L., Shi, W., Chen, R., Song, Z., Huang, C., Li, J., Dong, X., Zhou, Y., Liu, Q., Ma, L., Wang, X., Zhou, J., Liu, Y., Boja, E., Robles, A. I., Ma, W., Wang, P., ... Fan, J. (2019). Integrated Proteogenomic Characterization of HBV-Related Hepatocellular Carcinoma. *Cell*, *179*(2), 561–577.e22. https://doi.org/10.1016/j.cell.2019.08.052

Gaujoux, R., & Seoighe, C. (2010). A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*, *11*(1), 367. https://doi.org/10.1186/1471-2105-11-367

Gerlinger, M., Horswell, S., Larkin, J., Rowan, A. J., Salm, M. P., Varela, I., Fisher, R., McGranahan, N., Matthews, N., Santos, C. R., Martinez, P., Phillimore, B., Begum, S., Rabinowitz, A., Spencer-Dene, B., Gulati, S., Bates, P. A., Stamp, G., Pickering, L., ... Swanton, C. (2014). Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nature Genetics*, *46*(3), 225–+. https://doi.org/10.1038/ng.2891

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., Varela, I., Phillimore, B., Begum, S., McDonald, N. Q., Butler, A., Jones, D., Raine, K., Latimer, C., Santos, C. R., ... Swanton, C. (2012). Intratumor Heterogeneity and Branched Evolution Revealed by

Multiregion Sequencing. *New England Journal of Medicine*, *366*(10), 883–892. ://WOS:000301172500005

Gonzalez-Perez, A., & Lopez-Bigas, N. (2012). Functional impact bias reveals cancer drivers. *Nucleic Acids Research*, *40*(21), e169–e169. https://doi.org/10.1093/nar/gks743

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nat Rev Genet*, *17*(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Grasso, C. S., Tsoi, J., Onyshchenko, M., Abril-Rodriguez, G., Ross-Macdonald, P., Wind-Rotolo, M., Champhekar, A., Medina, E., Torrejon, D. Y., Shin, D. S., Tran, P., Kim, Y. J., Puig-Saus, C., Campbell, K., Vega-Crespo, A., Quist, M., Martignier, C., Luke, J. J., Wolchok, J. D., ... Ribas, A. (2020). Conserved Interferon-γ Signaling Drives Clinical Response to Immune Checkpoint Blockade Therapy in Melanoma. *Cancer Cell*. https://doi.org/10.1016/j.ccell.2020.08.005

Greaves, M. (2015). Evolutionary Determinants of Cancer. *Cancer Discovery*, *5*(8), 806–820. https://doi.org/10.1158/2159-8290.cd-15-0439

Gueguinou, M., Crottès, D., Chantôme, A., Rapetti-Mauss, R., Potier-Cartereau, M., Clarysse, L., Girault, A., Fourbon, Y., Jézéquel, P., Guérin-Charbonnel, C., Fromont, G., Martin, P., Pellissier, B., Schiappa, R., Chamorey, E., Mignen, O., Uguen, A., Borgese, F., Vandier, C., & Soriani, O. (2017). The SigmaR1 chaperone drives breast and colorectal cancer cell migration by tuning SK3-dependent Ca2+

homeostasis. *Oncogene*, *36*, 3640. https://doi.org/10.1038/onc.2016.501

Guichard, C., Amaddeo, G., Imbeaud, S., Ladeiro, Y., Pelletier, L., Maad, I. B., Calderaro, J., Bioulac-Sage, P., Letexier, M., Degos, F., Clement, B., Balabaud, C., Chevet, E., Laurent, A., Couchy, G., Letouze, E., Calvo, F., & Zucman-Rossi, J. (2012). Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. *Nat Genet*, *44*(6), 694–698. https://doi.org/10.1038/ng.2256

Hanahan, D., & Weinberg, R. A. (2000). The Hallmarks of Cancer. *Cell*, *100*(1), 57–70. https://doi.org/10.1016/S0092-8674(00)81683-9

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of Cancer: The Next Generation. *Cell*, *144*(5), 646–674. https://doi.org/10.1016/j.cell.2011.02.013

Hao, J.-J., Lin, D.-C., Dinh, H. Q., Mayakonda, A., Jiang, Y.-Y., Chang, C., Jiang, Y., Lu, C.-C., Shi, Z.-Z., Xu, X., Zhang, Y., Cai, Y., Wang, J.-W., Zhan, Q.-M., Wei, W.-Q., Berrnan, B. P., Wang, M.-R., & Koeffler, H. P. (2016). Spatial intratumoral heterogeneity and temporal clonal evolution in esophageal squamous cell carcinoma. *Nature Genetics*, *48*(12), 1500–1507. https://doi.org/10.1038/ng.3683

Harbst, K., Lauss, M., Cirenajwis, H., Isaksson, K., Rosengren, F., Torngren, T., Kvist, A., Johansson, M. C., Vallon-Christersson, J., Baldetorp, B., Borg, A., Olsson, H., Ingvar, C., Carneiro, A., & Jonsson, G. (2016). Multiregion Whole-Exome Sequencing

Uncovers the Genetic Evolution and Mutational Heterogeneity of Early-Stage Metastatic Melanoma. *Cancer Research*, *76*(16), 4765–4774. https://doi.org/10.1158/0008-5472.can-15-3476

Hayashi, A., Fan, J., Chen, R., Ho, Y., Makohon-Moore, A. P., Lecomte, N., Zhong, Y., Hong, J., Huang, J., Sakamoto, H., Attiyeh, M. A., Kohutek, Z. A., Zhang, L., Boumiza, A., Kappagantula, R., Baez, P., Bai, J., Lisi, M., Chadalavada, K., ... Iacobuzio-Donahue, C. A. (2020). A unifying paradigm for transcriptional heterogeneity and squamous features in pancreatic ductal adenocarcinoma. *Nature Cancer*, *1*(1), 59–74. https://doi.org/10.1038/s43018-019-0010-1

Hetz, C. (2012). The unfolded protein response: Controlling cell fate decisions under ER stress and beyond. *Nature Reviews Molecular Cell Biology*, *13*(2), 89–102. https://doi.org/10.1038/nrm3270

Hofree, M., Carter, H., Kreisberg, J. F., Bandyopadhyay, S., Mischel, P. S., Friend, S., & Ideker, T. (2016). Challenges in identifying cancer genes by analysis of exome sequencing data. *Nature Communications*, *7*, 12096. https://doi.org/10.1038/ncomms12096

Hoshida, Y., Brunet, J.-P., Tamayo, P., Golub, T. R., & Mesirov, J. P. (2007). Subclass Mapping: Identifying Common Subtypes in Independent Disease Data Sets. *PLoS ONE*, *2*(11). https://doi.org/10.1371/journal.pone.0001195

Hoshida, Y., Nijman, S. M. B., Kobayashi, M., Chan, J. A., Brunet, J.-P., Chiang, D. Y., Villanueva, A., Newell, P., Ikeda, K., Hashimoto, M.,

Watanabe, G., Gabriel, S., Friedman, S. L., Kumada, H., Llovet, J. M., & Golub, T. R. (2009). Integrative Transcriptome Analysis Reveals Common Molecular Subclasses of Human Hepatocellular Carcinoma. *Cancer Research*, *69*(18), 7385–7392. https://doi.org/10.1158/0008-5472.CAN-09-1089

Huang, J., Deng, Q., Wang, Q., Li, K.-Y., Dai, J.-H., Li, N., Zhu, Z.-D., Zhou, B., Liu, X.-Y., Liu, R.-F., Fei, Q.-L., Chen, H., Cai, B., Zhou, B., Xiao, H.-S., Qin, L.-X., & Han, Z.-G. (2012). Exome sequencing of hepatitis B virus-associated hepatocellular carcinoma. *Nat Genet*, *44*(10), 1117–1121. https://doi.org/abs/ng.2391.html#supplementary-information

Jamal-Hanjani, M., Quezada, S. A., Larkin, J., & Swanton, C. (2015). Translational Implications of Tumor Heterogeneity. *Clinical Cancer Research*, *21*(6), 1258. https://doi.org/10.1158/1078-0432.CCR-14-1429

Jamal-Hanjani, M., Wilson, G. A., McGranahan, N., Birkbak, N. J., Watkins, T. B. K., Veeriah, S., Shafi, S., Johnson, D. H., Mitter, R., Rosenthal, R., Salm, M., Horswell, S., Escudero, M., Matthews, N., Rowan, A., Chambers, T., Moore, D. A., Turajlic, S., Xu, H., ... Swanton, C. (2017). Tracking the Evolution of Non–Small-Cell Lung Cancer. *New England Journal of Medicine*, *376*(22), 2109–2121. https://doi.org/10.1056/NEJMoa1616288

Jhunjhunwala, S., Jiang, Z., Stawiski, E. W., Gnad, F., Liu, J., Mayba, O., Du, P., Diao, J., Johnson, S., Wong, K.-F., Gao, Z., Li, Y.,

Wu, T. D., Kapadia, S. B., Modrusan, Z., French, D. M., Luk, J. M., Seshagiri, S., & Zhang, Z. (2014). Diverse modes of genomic alteration in hepatocellular carcinoma. *Genome Biology*, *15*(8), 436. https://doi.org/10.1186/s13059-014-0436-9

Juul, M., Bertl, J., Guo, Q., Nielsen, M. M., Świtnicki, M., Hornshøj, H., Madsen, T., Hobolth, A., & Pedersen, J. S. (2017). Non-coding cancer driver candidates identified with a sample- and position-specific model of the somatic mutation rate. *eLife*, *6*, e21778. https://doi.org/10.7554/eLife.21778

Kamburov, A., Lawrence, M. S., Polak, P., Leshchiner, I., Lage, K., Golub, T. R., Lander, E. S., & Getz, G. (2015). Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proceedings of the National Academy of Sciences*, *112*(40), E5486. https://doi.org/10.1073/pnas.1516373112

Kamburov, A., Wierling, C., Lehrach, H., & Herwig, R. (2009). ConsensusPathDB—a database for integrating human functional interaction networks. *Nucleic Acids Research*, *37*(suppl_1), D623–D628. https://doi.org/10.1093/nar/gkn698

Kan, Z., Zheng, H., Liu, X., Li, S., Barber, T. D., Gong, Z., Gao, H., Hao, K., Willard, M. D., Xu, J., Hauptschein, R., Rejto, P. A., Fernandez, J., Wang, G., Zhang, Q., Wang, B., Chen, R., Wang, J., Lee, N. P., ... Mao, M. (2013). Whole-genome sequencing identifies recurrent mutations in hepatocellular carcinoma. *Genome Research*, *23*(9), 1422–1433. https:

//doi.org/10.1101/gr.154492.113

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, *28*(1), 27–30. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC102409/%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC102409/pdf/gkd027.pdf

Kasar, S., Kim, J., Improgo, R., Tiao, G., Polak, P., Haradhvala, N., Lawrence, M. S., Kiezun, A., Fernandes, S. M., Bahl, S., Sougnez, C., Gabriel, S., Lander, E. S., Kim, H. T., Getz, G., & Brown, J. R. (2015). Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nature Communications*, *6*, 8866. https://doi.org/10.1038/ncomms9866

Kim, R. D., Sarker, D., Meyer, T., Yau, T., Macarulla, T., Park, J.-W., Choo, S. P., Hollebecque, A., Sung, M. W., Lim, H.-Y., Mazzaferro, V., Trojan, J., Zhu, A. X., Yoon, J.-H., Sharma, S., Lin, Z.-Z., Chan, S. L., Faivre, S., Feun, L. G., ... Kang, Y.-K. (2019). First-in-Human Phase I Study of Fisogatinib (BLU-554) Validates Aberrant FGF19 Signaling as a Driver Event in Hepatocellular Carcinoma. *Cancer Discovery*, *9*(12), 1696–1707. https://doi.org/10.1158/2159-8290.CD-19-0555

Knudson, A. G. (1971). Mutation and Cancer: Statistical Study of Retinoblastoma. *Proceedings of the National Academy of Sciences*, *68*(4), 820–823. https://doi.org/10.1073/pnas.68.4.820

Lawlor, R. T., Veronese, N., Pea, A., Nottegar, A., Smith, L., Pilati, C., Demurtas, J., Fassan, M., Cheng, L., & Luchini, C. (2019). Alternative

lengthening of telomeres (ALT) influences survival in soft tissue sarcomas: A systematic review with meta-analysis. *BMC Cancer*, *19*(1), 232. https://doi.org/10.1186/s12885-019-5424-8

Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., Meyerson, M., Gabriel, S. B., Lander, E. S., & Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, *505*, 495. https://doi.org/10.1038/nature12912

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., Kiezun, A., Hammerman, P. S., McKenna, A., Drier, Y., Zou, L., Ramos, A. H., Pugh, T. J., Stransky, N., Helman, E., ... Getz, G. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, *499*, 214. https://doi.org/10.1038/nature12213

Lee, J.-S., Chu, I.-S., Heo, J., Calvisi, D. F., Sun, Z., Roskams, T., Durnez, A., Demetris, A. J., & Thorgeirsson, S. S. (2004). Classification and prediction of survival in hepatocellular carcinoma by gene expression profiling. *Hepatology*, *40*(3), 667–676. https://doi.org/10.1002/hep.20375

Lee, J.-S., Heo, J., Libbrecht, L., Chu, I.-S., Kaposi-Novak, P., Calvisi, D. F., Mikaelyan, A., Roberts, L. R., Demetris, A. J., Sun, Z., Nevens, F., Roskams, T., & Thorgeirsson, S. S. (2006). A novel prognostic subtype of human hepatocellular carcinoma derived from hepatic progenitor cells.

Nature Medicine, *12*(4), 410–416. https://doi.org/10.1038/nm1377

Letouzé, E., Shinde, J., Renault, V., Couchy, G., Blanc, J.-F., Tubacher, E., Bayard, Q., Bacq, D., Meyer, V., Semhoun, J., Bioulac-Sage, P., Prévôt, S., Azoulay, D., Paradis, V., Imbeaud, S., Deleuze, J.-F., & Zucman-Rossi, J. (2017). Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nature Communications*, *8*(1), 1315. https://doi.org/10.1038/s41467-017-01358-x

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, *26*(5), 589–595. https://doi.org/10.1093/bioinformatics/btp698

Li, Y. Y., Hanna, G. J., Laga, A. C., Haddad, R. I., Lorch, J. H., & Hammerman, P. S. (2015). Genomic analysis of metastatic cutaneous squamous cell carcinoma. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research*, *21*(6), 1447–1456. https://doi.org/10.1158/1078-0432.CCR-14-1773

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

Lin, D.-C., Mayakonda, A., Dinh, H. Q., Huang, P., Lin, L., Liu, X., Ding, L., Wang, J., Berman, B. P., Song, E.-W., Yin, D., & Koeffler, H. P. (2017). Genomic and Epigenomic Heterogeneity of Hepatocellular Carcinoma.

Cancer Research, *77*(9), 2255. https://doi.org/10.1158/0008-5472.CAN-16-2822

Ling, S., Hu, Z., Yang, Z., Yang, F., Li, Y., Lin, P., Chen, K., Dong, L., Cao, L., Tao, Y., Hao, L., Chen, Q., Gong, Q., Wu, D., Li, W., Zhao, W., Tian, X., Hao, C., Hungate, E. A., ... Wu, C.-I. (2015). Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution. *Proceedings of the National Academy of Sciences*, *112*(47), E6496–E6505. https://doi.org/10.1073/pnas.1519556112

Litovchick, L. (2011). DYRK1A protein kinase promotes quiescence and senescence through DREAM complex assembly. *Genes & Development*, *25*(8), 801–813.

Llovet, J. M., Ricci, S., Mazzaferro, V., Hilgard, P., Gane, E., Blanc, J.-F., Oliveira, A. C. de, Santoro, A., Raoul, J.-L., Forner, A., Schwartz, M., Porta, C., Zeuzem, S., Bolondi, L., Greten, T. F., Galle, P. R., Seitz, J.-F., Borbath, I., Häussinger, D., ... Bruix, J. (2008). Sorafenib in Advanced Hepatocellular Carcinoma. *New England Journal of Medicine*, *359*(4), 378–390. https://doi.org/10.1056/NEJMoa0708857

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Loveday, C., Litchfield, K., Proszek, P. Z., Cornish, A. J., Santo, F., Levy, M., Macintyre, G., Holryod, A., Broderick, P., Dudakia, D., Benton, B., Bakir, M. A., Hiley, C., Grist, E., Swanton,

C., Huddart, R., Powles, T., Chowdhury, S., Shipley, J., ... Turnbull, C. (2020). Genomic landscape of platinum resistant and sensitive testicular cancers. *Nature Communications*, *11*(1), 2189. https://doi.org/10.1038/s41467-020-15768-x

Luke, J. J., Bao, R., Sweis, R. F., Spranger, S., & Gajewski, T. F. (2019). WNT/β-catenin Pathway Activation Correlates with Immune Exclusion across Human Cancers. *Clinical Cancer Research*, *25*(10), 3074–3083. https://doi.org/10.1158/1078-0432.CCR-18-1942

Lupberger, J., & Hildt, E. (2007). Hepatitis B virus-induced oncogenesis. *World Journal of Gastroenterology : WJG*, *13*(1), 74–81. https://doi.org/10.3748/wjg.v13.i1.74

Lytle, N. K., Barber, A. G., & Reya, T. (2018). Stem cell fate in cancer growth, progression and therapy resistance. *Nature Reviews Cancer*, *18*(11), 669–680. https://doi.org/10.1038/s41568-018-0056-x

Maley, C. C., Aktipis, A., Graham, T. A., Sottoriva, A., Boddy, A. M., Janiszewska, M., Silva, A. S., Gerlinger, M., Yuan, Y., Pienta, K. J., Anderson, K. S., Gatenby, R., Swanton, C., Posada, D., Wu, C.-. I., Schiffman, J. D., Hwang, E. S., Polyak, K., Anderson, A. R. A., ... Shibata, D. (2017). Classifying the evolutionary and ecological features of neoplasms. *Nat Rev Cancer*, *17*(10), 605–619. https://doi.org/10.1038/nrc.2017.69

Manchester, K. L. (1995). Theodor Boveri and the origin of malignant tumours. *Trends in Cell Biology*, *5*(10), 384–387. https://doi.org/10.

1016/s0962-8924(00)89080-7

Mao, X., Yu, Y., Boyd, L. K., Ren, G., Lin, D., Chaplin, T., Kudahetti, S. C., Stankiewicz, E., Xue, L., Beltran, L., Gupta, M., Oliver, R. T. D., Lemoine, N. R., Berney, D. M., Young, B. D., & Lu, Y.-J. (2010). Distinct Genomic Alterations in Prostate Cancers in Chinese and Western Populations Suggest Alternative Pathways of Prostate Carcinogenesis. *Cancer Research*, *70*(13), 5207–5212. https://doi.org/10.1158/0008-5472.CAN-09-4074

Martincorena, I., Raine, K. M., Gerstung, M., Dawson, K. J., Haase, K., Van Loo, P., Davies, H., Stratton, M. R., & Campbell, P. J. (2018). Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell*, *171*(5), 1029–1041.e21. https://doi.org/10.1016/j.cell.2017.09.042

Marusyk, A., Almendro, V., & Polyak, K. (2012). Intra-tumour heterogeneity: A looking glass for cancer? *Nat Rev Cancer*, *12*(5), 323–334. https://doi.org/10.1038/nrc3261

Mäkinen, N., Aavikko, M., Heikkinen, T., Taipale, M., Taipale, J., Koivisto-Korander, R., Bützow, R., & Vahteristo, P. (2016). Exome Sequencing of Uterine Leiomyosarcomas Identifies Frequent Mutations in TP53, ATRX, and MED12. *PLOS Genetics*, *12*(2), e1005850. https://doi.org/10.1371/journal.pgen.1005850

McGranahan, N., Favero, F., Bruin, E. C. de, Birkbak, N. J., Szallasi, Z., & Swanton, C. (2015). Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Science*

Translational Medicine, *7*(283), 283ra54. https://doi.org/10.1126/scitranslmed.aaa1408

McGranahan, N., & Swanton, C. (2017). Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell*, *168*(4), 613–628. https://doi.org/10.1016/j.cell.2017.01.018

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., & Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, *12*(4), R41. https://doi.org/10.1186/gb-2011-12-4-r41

Metzker, M. L. (2010). Sequencing technologies [mdash] the next generation. *Nat Rev Genet*, *11*(1), 31–46. https://doi.org/10.1038/nrg2626

Mroz, E. A., Tward, A. D., Pickering, C. R., Myers, J. N., Ferris, R. L., & Rocco, J. W. (2013). High intratumor genetic heterogeneity is related to worse outcome in patients with head and neck squamous cell carcinoma. *Cancer*, *119*(16), 3034–3042. https://doi.org/10.1002/cncr.28150

Nault, J.-C., Paradis, V., Cherqui, D., Vilgrain, V., & Zucman-Rossi, J. (2017). Molecular classification of hepatocellular adenoma in clinical practice. *Journal of Hepatology*, *67*(5), 1074–1083. https://doi.org/10.1016/j.jhep.2017.07.009

Ng, A. W. T., Poon, S. L., Huang, M. N., Lim, J. Q., Boot, A., Yu, W., Suzuki, Y., Thangaraju, S., Ng, C. C. Y., Tan, P., Pang, S.-T., Huang,

H.-Y., Yu, M.-C., Lee, P.-H., Hsieh, S.-Y., Chang, A. Y., Teh, B. T., & Rozen, S. G. (2017). Aristolochic acids and their derivatives are widely implicated in liver cancers in Taiwan and throughout Asia. *Science Translational Medicine*, *9*(412). https://doi.org/10.1126/scitranslmed.aan6446

Nishimura, D. (2001). BioCarta. *Biotech Software & Internet Report*, *2*(3), 117–120. https://doi.org/10.1089/152791601750294344

Olivier, M., Hollstein, M., & Hainaut, P. (2010). TP53 Mutations in Human Cancers: Origins, Consequences, and Clinical Use. *Cold Spring Harbor Perspectives in Biology*, *2*(1), a001008–a001008. https://doi.org/10.1101/cshperspect.a001008

PCAWG Mutational Signatures Working Group, PCAWG Consortium, Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., Boot, A., Covington, K. R., Gordenin, D. A., Bergstrom, E. N., Islam, S. M. A., Lopez-Bigas, N., Klimczak, L. J., McPherson, J. R., Morganella, S., Sabarinathan, R., Wheeler, D. A., ... Stratton, M. R. (2020). The repertoire of mutational signatures in human cancer. *Nature*, *578*(7793), 94–101. https://doi.org/10.1038/s41586-020-1943-3

Prevarskaya, N., Skryma, R., & Shuba, Y. (2010). Ion channels and the hallmarks of cancer. *Trends in Molecular Medicine*, *16*(3), 107–121. https://doi.org/10.1016/j.molmed.2010.01.005

Ramos, A. H., Lichtenstein, L., Gupta, M., Lawrence, M. S., Pugh, T. J., Saksena, G., Meyerson, M., & Getz, G. (2015). Oncotator: Cancer

Variant Annotation Tool. *Human Mutation*, *36*(4), E2423–E2429. https://doi.org/10.1002/humu.22771

Razavi-Shearer, D., Gamkrelidze, I., Nguyen, M. H., Chen, D.-S., Damme, P. V., Abbas, Z., Abdulla, M., Rached, A. A., Adda, D., Aho, I., Akarca, U., Hasan, F., Lawati, F. A., Naamani, K. A., Al-Ashgar, H. I., Alavian, S. M., Alawadhi, S., Albillos, A., Al-Busafi, S. A., ... Razavi, H. (2018). Global prevalence, treatment, and prevention of hepatitis B virus infection in 2016: A modelling study. *The Lancet Gastroenterology & Hepatology*, *3*(6), 383–403. https://doi.org/10.1016/S2468-1253(18)30056-6

Reimand, J., Arak, T., Adler, P., Kolberg, L., Reisberg, S., Peterson, H., & Vilo, J. (2016). G:Profiler—a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Research*, *44*(W1), W83–W89. https://doi.org/10.1093/nar/gkw199

Rheinbay, E., Nielsen, M. M., Abascal, F., Wala, J. A., Shapira, O., Tiao, G., Hornshøj, H., Hess, J. M., Juul, R. I., Lin, Z., Feuerbach, L., Sabarinathan, R., Madsen, T., Kim, J., Mularoni, L., Shuai, S., Lanzós, A., Herrmann, C., Maruvka, Y. E., ... Getz, G. (2020). Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature*, *578*(7793), 102–111. https://doi.org/10.1038/s41586-020-1965-x

Rich, N. E., Hester, C., Odewole, M., Murphy, C. C., Parikh, N. D., Marrero, J. A., Yopp, A. C., & Singal, A. G. (2019). Racial and Ethnic Differences in Presentation and Outcomes of Hepatocellular Carcinoma. *Clinical*

Gastroenterology and Hepatology, *17*(3), 551–559.e1. https://doi.org/10.1016/j.cgh.2018.05.039

Ricketts, C. J., & Linehan, W. M. (2015). Gender Specific Mutation Incidence and Survival Associations in Clear Cell Renal Cell Carcinoma (CCRCC). *PLOS ONE*, *10*(10), e0140257. https://doi.org/10.1371/journal.pone.0140257

Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S., & Swanton, C. (2016). deconstructSigs: Delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biology*, *17*. https://doi.org/10.1186/s13059-016-0893-4

Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., Aparicio, S., Bouchard-Côté, A., & Shah, S. P. (2014). PyClone: Statistical inference of clonal population structure in cancer. *Nature Methods*, *11*, 396. https://doi.org/10.1038/nmeth.2883

Saunders, C. T., Wong, W. S. W., Swamy, S., Becq, J., Murray, L. J., & Cheetham, R. K. (2012). Strelka: Accurate somatic small-variant calling from sequenced tumor–normal sample pairs. *Bioinformatics*, *28*(14), 1811–1817. https://doi.org/10.1093/bioinformatics/bts271

Schulze, K., Imbeaud, S., Letouze, E., Alexandrov, L. B., Calderaro, J., Rebouissou, S., Couchy, G., Meiller, C., Shinde, J., Soysouvanh, F., Calatayud, A.-L., Pinyol, R., Pelletier, L., Balabaud, C., Laurent, A., Blanc, J.-F., Mazzaferro, V., Calvo, F., Villanueva, A., ...

Zucman-Rossi, J. (2015). Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat Genet*, *47*(5), 505–511. https://doi.org/10.1038/ng.3252

Shi, J.-Y., Xing, Q., Duan, M., Wang, Z.-C., Yang, L.-X., Zhao, Y.-J., Wang, X.-Y., Liu, Y., Deng, M., Ding, Z.-B., Ke, A.-W., Zhou, J., Fan, J., Cao, Y., Wang, J., Xi, R., & Gao, Q. (2016). Inferring the progression of multifocal liver cancer from spatial and temporal genomic heterogeneity. *Oncotarget*, *7*(3), 2867–2877. https://doi.org/10.18632/oncotarget.6558

Shibata, T., & Aburatani, H. (2014). Exploration of liver cancer genomes. *Nature Reviews Gastroenterology & Hepatology*, *11*(6), 340–349. https://doi.org/10.1038/nrgastro.2014.6

Sia, D., Jiao, Y., Martinez-Quetglas, I., Kuchuk, O., Villacorta-Martin, C., Castro de Moura, M., Putra, J., Camprecios, G., Bassaganyas, L., Akers, N., Losic, B., Waxman, S., Thung, S. N., Mazzaferro, V., Esteller, M., Friedman, S. L., Schwartz, M., Villanueva, A., & Llovet, J. M. (2017). Identification of an Immune-specific Class of Hepatocellular Carcinoma, Based on Molecular Features. *Gastroenterology*, *153*(3), 812–826. https://doi.org/10.1053/j.gastro.2017.06.007

Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., & Ng, P. C. (2012). SIFT web server: Predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research*, *40*(W1), W452–W457. https://doi.org/10.1093/nar/gks539

Stratton, M. R. (2011). Exploring the Genomes of Cancer Cells: Progress and Promise. *Science*, *331*(6024), 1553. https://doi.org/10.1126/science.1204040

Sun, L., Clavijo, P. E., Robbins, Y., Patel, P., Friedman, J., Greene, S., Das, R., Silvin, C., Van Waes, C., Horn, L. A., Schlom, J., Palena, C., Maeda, D., Zebala, J., & Allen, C. T. (2019). Inhibiting myeloid-derived suppressor cell trafficking enhances T cell immunotherapy. *JCI Insight*, *4*(7). https://doi.org/10.1172/jci.insight.126853

Tamborero, D., Gonzalez-Perez, A., & Lopez-Bigas, N. (2013). OncodriveCLUST: Exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*, *29*(18), 2238–2244. https://doi.org/10.1093/bioinformatics/btt395

Tan, D. S. W., Mok, T. S. K., & Rebbeck, T. R. (2015). Cancer Genomics: Diversity and Disparity Across Ethnicity and Geography. *Journal of Clinical Oncology*, *34*(1), 91–101. https://doi.org/10.1200/JCO.2015.62.0096

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., & Prins, P. (2015). Sambamba: Fast processing of NGS alignment formats. *Bioinformatics*, *31*(12), 2032–2034. https://doi.org/10.1093/bioinformatics/btv098

Tiwari, A., Mukherjee, B., Hassan, Md. K., Pattanaik, N., Jaiswal, A. M., & Dixit, M. (2019). Reduced FRG1 expression promotes prostate cancer progression and affects prostate cancer cell migration and invasion. *BMC Cancer*, *19*. https://doi.org/10.1186/s12885-019-5509-4

Tiwari, A., Pattnaik, N., Mohanty Jaiswal, A., & Dixit, M. (2017). Increased FSHD region gene1 expression reduces in vitro cell migration, invasion, and angiogenesis, ex vivo supported by reduced expression in tumors. *Bioscience Reports*, *37*(5). https://doi.org/10.1042/BSR20171062

Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B., & Karchin, R. (2016). Evaluating the evaluation of cancer driver genes. *Proc Natl Acad Sci U S A*, *113*(50), 14330–14335. https://doi.org/10.1073/pnas.1616440113

Totoki, Y., Tatsuno, K., Covington, K. R., Ueda, H., Creighton, C. J., Kato, M., Tsuji, S., Donehower, L. A., Slagle, B. L., Nakamura, H., Yamamoto, S., Shinbrot, E., Hama, N., Lehmkuhl, M., Hosoda, F., Arai, Y., Walker, K., Dahdouli, M., Gotoh, K., ... Shibata, T. (2014). Trans-ancestry mutational landscape of hepatocellular carcinoma genomes. *Nat Genet*, *46*(12), 1267–1273. https://doi.org/10.1038/ng.3126

Totoki, Y., Tatsuno, K., Yamamoto, S., Arai, Y., Hosoda, F., Ishikawa, S., Tsutsumi, S., Sonoda, K., Totsuka, H., Shirakihara, T., Sakamoto, H., Wang, L., Ojima, H., Shimada, K., Kosuge, T., Okusaka, T., Kato, K., Kusuda, J., Yoshida, T., ... Shibata, T. (2011). High-resolution characterization of a hepatocellular carcinoma genome. *Nat Genet*, *43*(5), 464–469. https://doi.org/10.1038/ng.804

Villanueva, A. (2019). Hepatocellular Carcinoma. *New England Journal of Medicine*, *380*(15), 1450–1462. https://doi.org/10.1056/NEJMra1713263

Virk, R. K. A., Wu, W., Almassalha, L. M., Bauer, G. M., Li, Y., VanDerway, D., Frederick, J., Zhang, D., Eshein, A., Roy, H. K., Szleifer, I., & Backman, V. (2020). Disordered chromatin packing regulates phenotypic plasticity. *Science Advances*, *6*(2), eaax6232. https://doi.org/10.1126/sciadv.aax6232

Wang, X.-M., Lu, Y., Song, Y.-M., Dong, J., Li, R.-Y., Wang, G.-L., Wang, X., Zhang, S.-D., Dong, Z.-H., Lu, M., Wang, S.-Y., Ge, L.-Y., Luo, G.-D., Ma, R.-Z., George Rozen, S., Bai, F., Wu, D., & Ma, L.-L. (2020). Integrative genomic study of Chinese clear cell renal cell carcinoma reveals features associated with thrombus. *Nature Communications*, *11*(1), 739. https://doi.org/10.1038/s41467-020-14601-9

Weber, R., Fleming, V., Hu, X., Nagibin, V., Groth, C., Altevogt, P., Utikal, J., & Umansky, V. (2018). Myeloid-Derived Suppressor Cells Hinder the Anti-Cancer Activity of Immune Checkpoint Inhibitors. *Frontiers in Immunology*, *9*, 1310. https://doi.org/10.3389/fimmu.2018.01310

Wilson, B. G., & Roberts, C. W. M. (2011). SWI/SNF nucleosome remodellers and cancer. *Nature Reviews Cancer*, *11*(7), 481–492. https://doi.org/10.1038/nrc3068

Wu, Y.-H., Graff, R. E., Passarelli, M. N., Hoffman, J. D., Ziv, E., Hoffmann, T. J., & Witte, J. S. (2018). Identification of Pleiotropic Cancer Susceptibility Variants from Genome-Wide Association Studies Reveals Functional Characteristics. *Cancer Epidemiology and Prevention Biomarkers*, *27*(1), 75–85. https://doi.org/10.1158/1055-9965.EPI-17-

0516

Xue, R., Li, R., Guo, H., Guo, L., Su, Z., Ni, X., Qi, L., Zhang, T., Li, Q., Zhang, Z., Xie, X. S., Bai, F., & Zhang, N. (2016). Variable Intra-Tumor Genomic Heterogeneity of Multiple Lesions in Patients With Hepatocellular Carcinoma. *Gastroenterology*, *150*(4), 998–1008. https://doi.org/10.1053/j.gastro.2015.12.033

Yates, L. R., Knappskog, S., Wedge, D., Farmery, J. H. R., Gonzalez, S., Martincorena, I., Alexandrov, L. B., Van Loo, P., Haugland, H. K., Lilleng, P. K., Gundem, G., Gerstung, M., Pappaemmanuil, E., Gazinska, P., Bhosle, S. G., Jones, D., Raine, K., Mudie, L., Latimer, C., ... Campbell, P. J. (2017). Genomic Evolution of Breast Cancer Metastasis and Relapse. *Cancer Cell*, *32*(2), 169–184.e7. https://doi.org/10.1016/j.ccell.2017.07.005

Yuan, J., Hu, Z., Mahal, B. A., Zhao, S. D., Kensler, K. H., Pi, J., Hu, X., Zhang, Y., Wang, Y., Jiang, J., Li, C., Zhong, X., Montone, K. T., Guan, G., Tanyi, J. L., Fan, Y., Xu, X., Morgan, M. A., Long, M., ... Zhang, L. (2018). Integrated Analysis of Genetic Ancestry and Genomic Alterations across Cancers. *Cancer Cell*, *34*(4), 549–560.e9. https://doi.org/10.1016/j.ccell.2018.08.019

Yue, P., Forrest, W. F., Kaminker, J. S., Lohr, S., Zhang, Z., & Cavet, G. (2010). Inferring the functional effects of mutation through clusters of mutations in homologous proteins. *Human Mutation*, *31*(3), 264–271. https://doi.org/10.1002/humu.21194

Zhai, W., Lim, T. K.-H., Zhang, T., Phang, S.-T., Tiang, Z., Guan, P., Ng, M.-H., Lim, J. Q., Yao, F., Li, Z., Ng, P. Y., Yan, J., Goh, B. K., Chung, A. Y.-F., Choo, S.-P., Khor, C. C., Soon, W. W.-J., Sung, K. W.-K., Foo, R. S.-Y., & Chow, P. K.-H. (2017). The spatial organization of intra-tumour heterogeneity and evolutionary trajectories of metastases in hepatocellular carcinoma. *Nature Communications*, *8*, 4565. https://doi.org/10.1038/ncomms14565

Zhang, Junjun, Baran, J., Cros, A., Guberman, J. M., Haider, S., Hsu, J., Liang, Y., Rivkin, E., Wang, J., Whitty, B., Wong-Erasmus, M., Yao, L., & Kasprzyk, A. (2011). International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data. *Database*, *2011*, bar026–bar026. https://doi.org/10.1093/database/bar026

Zhang, Jianjun, Fujimoto, J., Zhang, J., Wedge, D. C., Song, X., Zhang, J., Seth, S., Chow, C.-W., Cao, Y., Gumbs, C., Gold, K. A., Kalhor, N., Little, L., Mahadeshwar, H., Moran, C., Protopopov, A., Sun, H., Tang, J., Wu, X., ... Futreal, A. (2014). Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science*, *346*(6206), 256–259. https://doi.org/10.1126/science.1256930

Zhang, W., He, H., Zang, M., Wu, Q., Zhao, H., Lu, L., Ma, P., Zheng, H., Wang, N., Zhang, Y., He, S., Chen, X., Wu, Z., Wang, X., Cai, J., Liu, Z., Sun, Z., Zeng, Y., Qu, C., & Jiao, Y. (2017). Genetic Features of Aflatoxin-associated Hepatocellular Carcinomas. *Gastroenterology*. https://doi.org/10.1053/j.gastro.2017.03.024

Zucman-Rossi, J., & Nault, J.-C. (2020). Mutations and Genomic Alterations in Liver Cancer. In *The Liver* (pp. 773–781). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119436812.ch60

Appendix

Table S1: Driver genes from MutSigCV.

gene	p	q	Frequency
TP53	0.0000000	0.0000000	0.2868790
CTNNB1	0.0000000	0.0000000	0.2831730
APOB	0.0000004	0.0001723	0.1074870
ARID1A	0.0000000	0.0000000	0.0919199
ALB	0.0000000	0.0000000	0.0889548
ARID2	0.0000000	0.0000000	0.0733877
AXIN1	0.0000000	0.0000000	0.0674574
COL11A1	0.0000094	0.0036358	0.0622683
NFE2L2	0.0000000	0.0000000	0.0437361
RPS6KA3	0.0000000	0.0000000	0.0429948
TSC2	0.0000000	0.0000000	0.0422535
DOCK2	0.0000001	0.0000247	0.0422535
RB1	0.0000000	0.0000000	0.0407709
SETD2	0.0000293	0.0110381	0.0348406
ACVR2A	0.0000000	0.0000000	0.0348406
KEAP1	0.0000000	0.0000000	0.0340993
BAP1	0.0000000	0.0000000	0.0252039
CDKN2A	0.0000000	0.0000000	0.0229800
SF3B1	0.0000959	0.0306593	0.0222387
NEFH	0.0000627	0.0219082	0.0222387
VAV3	0.0000793	0.0265394	0.0214974
BRD7	0.0000000	0.0000000	0.0214974
PTEN	0.0000000	0.0000000	0.0207561
HNF1A	0.0000000	0.0000000	0.0207561
IL6ST	0.0000000	0.0000000	0.0200148
TF	0.0000000	0.0000000	0.0192735
TSC1	0.0000000	0.0000127	0.0177910
DYRK1A	0.0000012	0.0004735	0.0170497
HNF4A	0.0000000	0.0000000	0.0163084
CDKN1A	0.0000000	0.0000000	0.0155671
CYP2E1	0.0000000	0.0000000	0.0148258

KCNN3	0.0000898	0.0292168	0.0140845
ERRFI1	0.0000511	0.0181727	0.0140845
RPL22	0.0000000	0.0000000	0.0133432
HNRNPA2B1	0.0000318	0.0117623	0.0133432
SELPLG	0.0000000	0.0000000	0.0126019
SLC30A1	0.0000000	0.0000000	0.0118606
ZFP36L1	0.0000000	0.0000000	0.0111193
KRAS	0.0000987	0.0310354	0.0111193
FRG1	0.0000000	0.0000000	0.0111193
ADH1B	0.0000779	0.0265394	0.0111193
PHF10	0.0001121	0.0346589	0.0103781
HP	0.0000802	0.0265394	0.0103781
CRIP3	0.0000000	0.0000000	0.0103781
CELF1	0.0000000	0.0000000	0.0103781

Table S2: Driver genes from TUSON Explorer

Gene	TSG q-value	OG q-value	Frequency
ACVR2A	0.0006598	1.0000000	0.0348406
ALB	0.0000000	1.0000000	0.0889548
ARID1B	0.0000000	1.0000000	0.0200148
AXIN1	0.0000000	1.0000000	0.0674574
BAP1	0.0008000	1.0000000	0.0252039
BRD7	0.0001181	1.0000000	0.0214974
CDKN2A	0.0060115	1.0000000	0.0229800
CTNNB1	1.0000000	0.0000118	0.2831730
DYRK1A	0.0175904	1.0000000	0.0170497
NFE2L2	1.0000000	0.0259487	0.0437361
PTEN	0.0353245	1.0000000	0.0207561
RB1	0.0000000	1.0000000	0.0407709
RPS6KA3	0.0000000	1.0000000	0.0429948
TP53	0.0000000	1.0000000	0.2868790
TSC1	0.0721570	1.0000000	0.0177910
TSC2	0.0000022	1.0000000	0.0422535

Table S3: Driver genes from 20/20+

Gene	oncogene q-value	TSG q-value	driver q-value	Frequency
CTNNB1	0.0000000	1.0000000	0.0000000	0.2831730
TSC1	0.9978700	0.0076923	0.0052632	0.0177910
TSC2	0.9995138	0.0076923	0.0052632	0.0422535
SETD2	0.9978700	0.0076923	0.0052632	0.0348406
TP53	0.8996612	0.0380952	0.0052632	0.2868790
BAP1	0.9978700	0.0076923	0.0052632	0.0252039
PIK3CA	0.0000000	1.0000000	0.0052632	0.0244626
KMT2B	0.9978700	0.0076923	0.0052632	0.0214974
BRD7	0.9978700	0.0076923	0.0052632	0.0214974
ARID1A	0.9978700	0.0076923	0.0052632	0.0919199
RB1	0.9978700	0.0076923	0.0052632	0.0407709
IDH1	0.0000000	1.0000000	0.0052632	0.0111193
RPS6KA3	0.9978700	0.0076923	0.0052632	0.0429948
NFE2L2	0.0000000	1.0000000	0.0052632	0.0437361
AXIN1	0.9978700	0.0076923	0.0052632	0.0674574
DYRK1A	0.9978700	0.0076923	0.0052632	0.0170497
ARID2	0.9978700	0.0076923	0.0052632	0.0733877
KRAS	0.0000000	1.0000000	0.0052632	0.0111193
CDKN2A	0.9943211	0.0235294	0.0100000	0.0229800
APC	0.9139812	0.0535714	0.0160000	0.0303929
KDM6A	0.9978700	0.0214286	0.0160000	0.0126019
PBRM1	0.9978700	0.0277778	0.0160000	0.0192735
KMT2D	0.9978700	0.0235294	0.0160000	0.0474426
CYP2E1	0.9978700	0.0368421	0.0230769	0.0148258
ATRX	0.9977330	0.0440000	0.0296296	0.0200148
TLE1	0.9800859	0.0617647	0.0303030	0.0126019
EEF1A1	0.0000000	1.0000000	0.0303030	0.0207561
ALB	0.9978700	0.0440000	0.0303030	0.0889548
CDKN1A	0.9978700	0.0380952	0.0303030	0.0155671
HNF1A	0.9978700	0.0500000	0.0382353	0.0207561
GSE1	0.9978700	0.0440000	0.0428571	0.0148258
SRCAP	0.9902035	0.0617647	0.0513514	0.0296516
TNRC6B	0.9978700	0.0535714	0.0513514	0.0259451

RPL22	0.9978700	0.0617647	0.0547619	0.0133432
SF3B1	0.0000000	0.9993911	0.0702128	0.0222387
PTPN3	0.9978700	0.0763158	0.0734694	0.0155671
NCOR1	0.8659292	0.1939394	0.0734694	0.0244626
RAPGEF2	0.9843679	0.1173077	0.0800000	0.0103781
ARID1B	0.9978700	0.1173077	0.0982759	0.0200148

Table S4: Driver gene list collected from the literature

Driver	Cleary(2013)	Kan (2013)	Totoki(2014)	Ahn (2014)	Schulze (2015)	Fujimato (2016)	TCGA (2017)	Chaudhary (2019)
TP53	1	1	1	1	1	1	1	1
CTNNB1	1	1	1	1	1	1	1	1
AXIN1	0	1	1	1	1	1	1	1
RPS6KA3	0	0	1	1	1	1	1	1
RB1	0	0	1	1	1	1	1	1
NFE2L2	0	0	1	0	1	1	1	1
CDKN2A	0	0	1	0	1	1	1	1
ARID1A	0	0	1	0	1	1	1	1
ARID2	0	0	1	0	1	1	1	0
ACVR2A	0	0	1	0	1	1	1	1
ALB	0	0	0	1	1	1	1	1
KEAP1	1	0	0	0	1	0	1	0
BRD7	0	0	1	0	0	1	0	0
CCND1	0	0	1	0	0	1	0	0
PTEN	0	0	1	0	0	1	0	0
CDKN1A	0	0	1	0	1	0	0	0
APOB	0	0	0	0	0	1	1	0
G6PC	0	0	1	0	0	1	0	0
RPL22	0	0	0	0	1	1	0	0
LRP1B	0	1	0	0	0	1	0	0
BAP1	0	0	0	0	0	1	1	0
COL11A1	0	1	0	0	0	0	0	0

AHCTF1	0	0	0	0	0	0	1	0
MEN1	0	0	1	0	0	0	0	0
IL6ST	0	0	0	0	0	0	1	0
RP1L1	0	0	0	0	0	0	1	0
SETDB1	0	0	0	0	0	1	0	0
TMEM51	1	0	0	0	0	0	0	0
CDKN2B	0	0	1	0	0	0	0	0
PER3	0	0	0	0	0	1	0	0
EYS	0	0	0	0	0	1	0	0
BRD9	1	0	0	0	0	0	0	0
MUC17	0	0	0	0	0	1	0	0
GPATCH4	0	0	0	0	0	0	1	0
ERGIC1	0	0	0	0	0	1	0	0
CACNA2D4	0	1	0	0	0	0	0	0
KRAS	0	0	0	0	0	0	1	0
ADRA1A	0	0	0	0	0	1	0	0
ADCY2	0	1	0	0	0	0	0	0
WDTC1	0	0	0	0	0	1	0	0
LZTR1	0	0	0	0	0	0	1	0
CPS1	0	0	0	0	0	1	0	0
CDKN1B	0	0	0	1	0	0	0	0
CPA2	1	0	0	0	0	0	0	0
JAK1	0	1	0	0	0	0	0	0
TSC1	0	0	1	0	0	0	0	0
TMEM170A	1	0	0	0	0	0	0	0

TMEM99	0	0	1	0	0	0	0	0
GJA1	1	0	0	0	0	0	0	0
HNF4A	0	0	0	0	0	1	0	0
CYP2E1	0	0	1	0	0	0	0	0
PIK3CA	0	0	0	0	0	0	1	0
FAM5C	0	1	0	0	0	0	0	0
TSC2	0	0	1	0	0	0	0	0
NCOR1	0	0	1	0	0	0	0	0
KRTAP5-11	0	0	0	0	0	1	0	0
SLC10A1	0	1	0	0	0	0	0	0
FGF19	0	0	1	0	0	0	0	0
TTC28	0	0	0	0	0	1	0	0
ASH1L	0	0	0	0	0	1	0	0
HIST1H1C	0	0	0	0	0	0	1	0
MAP2K3	0	0	1	0	0	0	0	0
TBL1XR1	0	0	0	0	0	1	0	0
GALNT11	0	0	1	0	0	0	0	0
PRKACA	0	0	0	0	0	1	0	0
NSMCE2	0	0	0	0	0	1	0	0
SRCAP	0	0	1	0	0	0	0	0
EEF1A1	0	0	0	0	0	0	1	0
SF3B1	0	0	0	0	0	0	1	0
IGSF3	1	0	0	0	0	0	0	0
SMARCA4	0	0	0	0	0	0	1	0
EPS15	0	1	0	0	0	0	0	0

NRAS	0	0	0	0	0	0	1	0
SELPLG	0	0	0	1	0	0	0	0
HNRNPA2B1	0	0	1	0	0	0	0	0
FCRL1	0	0	1	0	0	0	0	0
TERT	0	0	1	0	0	0	0	0
ESRRG	0	0	0	0	0	1	0	0
PCMTD1	1	0	0	0	0	0	0	0
ATAD3B	1	0	0	0	0	0	0	0
TTLL2	1	0	0	0	0	0	0	0
AR	1	0	0	0	0	0	0	0
VCX	0	0	0	1	0	0	0	0
MTAP	0	0	0	0	0	1	0	0
AZIN1	0	0	0	0	0	0	1	0
CREB3L3	0	0	0	0	0	0	1	0
ADH1B	0	0	1	0	0	0	0	0
MACROD2	0	0	0	0	0	1	0	0

Table S5: Final driver gene list and novelty status

Gene	TUSON	MutSigCV	2020plus	CGC	Reported	Status
APC	0	0	1	1	0	Novel
ARID1B	1	0	1	1	0	Novel
ATRX	0	0	1	1	0	Novel
CELF1	0	1	0	0	0	Novel
CRIP3	0	1	0	0	0	Novel
DOCK2	0	1	0	0	0	Novel
DYRK1A	1	1	1	0	0	Novel
ERRFI1	0	1	0	0	0	Novel
FRG1	0	1	0	0	0	Novel
GSE1	0	0	1	0	0	Novel
HNF1A	0	1	1	1	0	Novel
HP	0	1	0	0	0	Novel
IDH1	0	0	1	1	0	Novel
KCNN3	0	1	0	0	0	Novel
KDM6A	0	0	1	1	0	Novel
KMT2B	0	0	1	0	0	Novel
KMT2D	0	0	1	1	0	Novel
NEFH	0	1	0	0	0	Novel
PBRM1	0	0	1	1	0	Novel
PHF10	0	1	0	0	0	Novel
PTPN3	0	0	1	0	0	Novel
RAPGEF2	0	0	1	0	0	Novel
SETD2	0	1	1	1	0	Novel
SLC30A1	0	1	0	0	0	Novel
TF	0	1	0	0	0	Novel
TLE1	0	0	1	0	0	Novel
TNRC6B	0	0	1	0	0	Novel
VAV3	0	1	0	0	0	Novel
ZFP36L1	0	1	0	0	0	Novel
ACVR2A	1	1	0	1	1	Reported
ADH1B	0	1	0	0	1	Reported
ALB	1	1	1	0	1	Reported
APOB	0	1	0	0	1	Reported

ARID1A	0	1	1	1	1	Reported
ARID2	0	1	1	1	1	Reported
AXIN1	1	1	1	1	1	Reported
BAP1	1	1	1	1	1	Reported
BRD7	1	1	1	0	1	Reported
CDKN1A	0	1	1	1	1	Reported
CDKN2A	1	1	1	1	1	Reported
COL11A1	0	1	0	0	1	Reported
CTNNB1	1	1	1	1	1	Reported
CYP2E1	0	1	1	0	1	Reported
EEF1A1	0	0	1	0	1	Reported
HNF4A	0	1	0	0	1	Reported
HNRNPA2B1	0	1	0	1	1	Reported
IL6ST	0	1	0	1	1	Reported
KEAP1	0	1	0	1	1	Reported
KRAS	0	1	1	1	1	Reported
NCOR1	0	0	1	1	1	Reported
NFE2L2	1	1	1	1	1	Reported
PIK3CA	0	0	1	1	1	Reported
PTEN	1	1	0	1	1	Reported
RB1	1	1	1	1	1	Reported
RPL22	0	1	1	1	1	Reported
RPS6KA3	1	1	1	0	1	Reported
SELPLG	0	1	0	0	1	Reported
SF3B1	0	1	1	1	1	Reported
SRCAP	0	0	1	0	1	Reported
TP53	1	1	1	1	1	Reported
TSC1	1	1	1	1	1	Reported
TSC2	1	1	1	1	1	Reported

Table S6: Identified oncogene and tumor suppressor genes

Gene	q-value	Type
ALB	0.0440000	Tumor suppressor
APC	0.0535714	Tumor suppressor

ARID1A	0.0076923	Tumor suppressor
ARID2	0.0076923	Tumor suppressor
ATRX	0.0440000	Tumor suppressor
AXIN1	0.0076923	Tumor suppressor
BAP1	0.0076923	Tumor suppressor
BRD7	0.0076923	Tumor suppressor
CDKN1A	0.0380952	Tumor suppressor
CDKN2A	0.0235294	Tumor suppressor
CYP2E1	0.0368421	Tumor suppressor
DYRK1A	0.0076923	Tumor suppressor
GSE1	0.0440000	Tumor suppressor
HNF1A	0.0500000	Tumor suppressor
KDM6A	0.0214286	Tumor suppressor
KMT2B	0.0076923	Tumor suppressor
KMT2D	0.0235294	Tumor suppressor
PBRM1	0.0277778	Tumor suppressor
PTPN3	0.0763158	Tumor suppressor
RB1	0.0076923	Tumor suppressor
RPL22	0.0617647	Tumor suppressor
RPS6KA3	0.0076923	Tumor suppressor
SETD2	0.0076923	Tumor suppressor
SRCAP	0.0617647	Tumor suppressor
TLE1	0.0617647	Tumor suppressor
TNRC6B	0.0535714	Tumor suppressor
TP53	0.0380952	Tumor suppressor
TSC1	0.0076923	Tumor suppressor
TSC2	0.0076923	Tumor suppressor
PTEN	0.0353245	Tumor suppressor
CRIP3	0.0000000	Oncogene
CTNNB1	0.0000000	Oncogene
EEF1A1	0.0000000	Oncogene
IDH1	0.0000000	Oncogene
KCNN3	0.0129032	Oncogene
KRAS	0.0000000	Oncogene
NFE2L2	0.0000000	Oncogene
PIK3CA	0.0000000	Oncogene

SF3B1	0.0000000	Oncogene

Table S7: dN/dS q-values for individual genes and the type of significance

Gene_name	Missense q-value	Truncating q-value	dN/dS Significance
TP53	0.0000000	0.0000000	Both
RPS6KA3	0.0000900	0.0000000	Both
ARID1A	0.0321074	0.0000000	Both
CDKN2A.p16INK4a	0.0000292	0.0000000	Both
NFE2L2	0.0000000	0.0864780	Both
ALB	0.0000974	0.0000094	Both
CDKN2A.p14arf	0.0000549	0.0000273	Both
RPL22	0.0021173	0.0000030	Both
CDKN1A	0.0014594	0.0000089	Both
CRIP3	0.0000072	0.0015344	Both
ACVR2A	0.0005125	0.0000888	Both
FRG1	0.0012677	0.0000049	Both
DYRK1A	0.0832193	0.0000056	Both
KEAP1	0.0000629	0.0318222	Both
HNF1A	0.0300701	0.0014195	Both
COL11A1	0.0003662	0.0335744	Both
PTEN	0.0195975	0.0144533	Both
VAV3	0.0017920	0.0037617	Both
DOCK2	0.0575294	0.0012272	Both
HNF4A	0.0362468	0.0178121	Both
ADH1B	0.0510690	0.0273657	Both
HP	0.0898203	0.0367188	Both

APOB	0.0716054	0.0865141	Both
CTNNB1	0.0000000	0.4702301	Missense
KRAS	0.0000111	0.6219472	Missense
PIK3CA	0.0001551	0.3044737	Missense
EEF1A1	0.0001030	0.4239838	Missense
HNRNPA2B1	0.0195975	0.4397175	Missense
IDH1	0.0033190	0.5434671	Missense
SF3B1	0.0042144	0.7693519	Missense
TF	0.0309252	0.5434671	Missense
SLC30A1	0.0510690	0.5786707	Missense
ARID2	0.2668274	0.0000000	Truncated
AXIN1	0.2669858	0.0000000	Truncated
RB1	0.4133731	0.0000000	Truncated
BRD7	0.8277898	0.0000000	Truncated
TSC2	0.2306349	0.0000000	Truncated
TSC1	0.5882211	0.0000001	Truncated
TNRC6B	0.5549171	0.0000540	Truncated
BAP1	0.2668274	0.0000094	Truncated
ERRFI1	0.7266804	0.0009209	Truncated
SETD2	0.7266804	0.0009185	Truncated
KDM6A	0.7266804	0.0019892	Truncated
CYP2E1	0.2480747	0.0013522	Truncated
TLE1	0.8930789	0.0019892	Truncated
APC	0.7811472	0.0037617	Truncated
KMT2D	0.2267424	0.0012272	Truncated

PHF10	0.7266804	0.0064466	Truncated
GSE1	0.1505489	0.0367188	Truncated
PTPN3	0.2008839	0.0032777	Truncated
CELF1	0.1605673	0.0032777	Truncated
RAPGEF2	0.3538579	0.0367188	Truncated
KMT2B	0.8277898	0.0093477	Truncated
PBRM1	0.9722983	0.0175209	Truncated
ARID1B	0.5863830	0.0312675	Truncated
SRCAP	0.5863830	0.0144533	Truncated
ATRX	0.8930789	0.0318408	Truncated

Table S8: Signature proportions across TCGA patients

Patient_ID	SBS1	SBS4	SBS5	SBS6	SBS12	SBS16	SBS22	SBS23	SBS24
TCGA-2V-A95S	0.0688455	0.0839475	0.7535272	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-2Y-A9GS	0.0000000	0.2628370	0.1587599	0.2409094	0.1352906	0.0000000	0.0000000	0.0000000	0.1517508
TCGA-2Y-A9GT	0.0629393	0.0943117	0.5317419	0.0000000	0.1539699	0.0000000	0.0000000	0.0000000	0.1112014
TCGA-2Y-A9GU	0.1033084	0.1213315	0.4834468	0.0000000	0.1480350	0.0000000	0.0666141	0.0772643	0.0000000
TCGA-2Y-A9GV	0.0000000	0.1578941	0.5110626	0.0000000	0.0697182	0.1171812	0.1441440	0.0000000	0.0000000
TCGA-2Y-A9GW	0.1011250	0.0785670	0.3768956	0.0000000	0.1843116	0.0000000	0.0791446	0.1017504	0.0782058
TCGA-2Y-A9GX	0.0000000	0.1148118	0.8209736	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-2Y-A9GY	0.0000000	0.1694171	0.6730534	0.0000000	0.0000000	0.0883035	0.0000000	0.0000000	0.0000000
TCGA-2Y-A9GZ	0.0000000	0.1243169	0.7636573	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0720904
TCGA-2Y-A9H0	0.0000000	0.0000000	0.7794496	0.0000000	0.0000000	0.0000000	0.0790715	0.0000000	0.0940137
TCGA-2Y-A9H1	0.0000000	0.1163234	0.5486958	0.0000000	0.1315118	0.0000000	0.0000000	0.0639596	0.0000000
TCGA-2Y-A9H2	0.2204106	0.0903481	0.6715758	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-2Y-A9H3	0.0000000	0.1323795	0.7533011	0.0000000	0.0670776	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-2Y-A9H4	0.0000000	0.2003929	0.5902976	0.0000000	0.0674002	0.0000000	0.0000000	0.1049616	0.0000000
TCGA-2Y-A9H5	0.1516955	0.0000000	0.7325952	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-2Y-A9H6	0.1239937	0.0000000	0.5946144	0.0000000	0.0000000	0.1190119	0.0000000	0.0000000	0.0681960
TCGA-2Y-A9H7	0.1092080	0.1105379	0.6784056	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-2Y-A9H8	0.2459872	0.1399817	0.4207378	0.0000000	0.0854227	0.0000000	0.0000000	0.1078706	0.0000000
TCGA-2Y-A9H9	0.1062292	0.2596343	0.4049077	0.0000000	0.1279046	0.0805779	0.0000000	0.0000000	0.0000000
TCGA-2Y-A9HA	0.0000000	0.2184705	0.4168605	0.0000000	0.0847342	0.0997957	0.0000000	0.0000000	0.0942797
TCGA-2Y-A9HB	0.0000000	0.1964008	0.4998383	0.1179168	0.1688547	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-3K-AAZ8	0.0000000	0.0000000	0.7431422	0.0000000	0.0000000	0.0000000	0.0838201	0.0000000	0.0000000

TCGA-4R-AA8I	0.0869646	0.0638786	0.0000000	0.4437345	0.3208847	0.0000000	0.0000000	0.0000000	0.0692741
TCGA-5C-A9VG	0.1027518	0.1898920	0.6022931	0.0000000	0.0000000	0.0000000	0.0000000	0.0751809	0.0000000
TCGA-5C-A9VH	0.0000000	0.2165367	0.6412771	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-5C-AAPD	0.0727218	0.2583480	0.5112273	0.0000000	0.0846638	0.0000000	0.0000000	0.0730391	0.0000000
TCGA-5R-AA1C	0.0000000	0.1249456	0.6660590	0.0000000	0.0000000	0.0954176	0.1135778	0.0000000	0.0000000
TCGA-5R-AA1D	0.1604588	0.0000000	0.7267843	0.0000000	0.0000000	0.0000000	0.0836416	0.0000000	0.0000000
TCGA-5R-AAAM	0.0000000	0.0000000	0.5562214	0.1213366	0.1671144	0.0636400	0.0000000	0.0000000	0.0000000
TCGA-BC-A10R	0.0000000	0.1646176	0.7131926	0.0647345	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-BC-A10S	0.1500141	0.0000000	0.7300782	0.0000000	0.1059998	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-BC-A10T	0.0000000	0.1918080	0.6177445	0.0000000	0.0676761	0.0000000	0.0714571	0.0000000	0.0000000
TCGA-BC-A10U	0.0000000	0.2095086	0.7133591	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-BC-A10W	0.0000000	0.1819025	0.7168427	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-BC-A10X	0.0000000	0.3523078	0.0000000	0.0710989	0.3834326	0.0000000	0.0000000	0.1496606	0.0000000
TCGA-BC-A10Y	0.0000000	0.1423044	0.4149000	0.1906236	0.1039013	0.1246257	0.0000000	0.0000000	0.0000000
TCGA-BC-A10Z	0.0000000	0.2100818	0.6222560	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.1072468
TCGA-BC-A110	0.2421990	0.0748855	0.6582435	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-BC-A216	0.1159166	0.0000000	0.8112540	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-BC-A217	0.0000000	0.2510156	0.5702370	0.0000000	0.0764338	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-BC-A3KF	0.0000000	0.0608587	0.4913512	0.0000000	0.1780019	0.0000000	0.1277504	0.0000000	0.1253568
TCGA-BC-A5W4	0.1149374	0.1110383	0.6051595	0.0000000	0.0744468	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-BC-A69H	0.0000000	0.1895699	0.6200082	0.0935850	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-BC-A69I	0.1363800	0.0000000	0.7198813	0.0000000	0.1379535	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-BC-A8YO	0.0933208	0.1768498	0.5243613	0.0000000	0.0725958	0.0000000	0.1193702	0.0000000	0.0000000
TCGA-BD-A2L6	0.0000000	0.1937037	0.6734535	0.0000000	0.0000000	0.0000000	0.0634735	0.0000000	0.0000000
TCGA-BD-A3EP	0.0000000	0.1527401	0.7338226	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000

TCGA-BD-A3ER	0.0000000	0.1696831	0.6705546	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.1173078
TCGA-BW-A5NO	0.1103935	0.1246754	0.6344963	0.0000000	0.0000000	0.0000000	0.0000000	0.0705450	0.0000000
TCGA-BW-A5NP	0.1424400	0.1610810	0.6018697	0.0000000	0.0000000	0.0000000	0.0000000	0.0946093	0.0000000
TCGA-BW-A5NQ	0.0000000	0.1508885	0.7891049	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-CC-5258	0.0000000	0.0000000	0.4394784	0.0782971	0.0831304	0.0657766	0.0000000	0.0000000	0.2815580
TCGA-CC-5259	0.0613393	0.1704692	0.5123651	0.0000000	0.0000000	0.0000000	0.2463579	0.0000000	0.0000000
TCGA-CC-5260	0.2144138	0.0000000	0.6069560	0.0000000	0.0000000	0.0000000	0.1786303	0.0000000	0.0000000
TCGA-CC-5261	0.2692924	0.0000000	0.6552823	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-CC-5262	0.0000000	0.1823705	0.4788676	0.0000000	0.0000000	0.0688403	0.0787460	0.0000000	0.0793955
TCGA-CC-5263	0.0000000	0.3378871	0.3285046	0.0000000	0.0861463	0.0000000	0.1558928	0.0000000	0.0000000
TCGA-CC-5264	0.0000000	0.2131336	0.5394883	0.0000000	0.0702011	0.0000000	0.1018884	0.0000000	0.0000000
TCGA-CC-A123	0.0000000	0.3217344	0.5848575	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-CC-A1HT	0.0000000	0.2356399	0.6067328	0.0000000	0.0000000	0.0606122	0.0000000	0.0000000	0.0000000
TCGA-CC-A3M9	0.0000000	0.1164224	0.4855726	0.0000000	0.0984516	0.0889349	0.1194823	0.0000000	0.0000000
TCGA-CC-A3MA	0.1866963	0.1181022	0.6250516	0.0000000	0.0701500	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-CC-A3MB	0.0807051	0.2095711	0.5233912	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.1182935
TCGA-CC-A3MC	0.0000000	0.1819536	0.3743592	0.0000000	0.0000000	0.1011008	0.0992151	0.0000000	0.1420473
TCGA-CC-A5UC	0.1249776	0.1491680	0.5082991	0.0000000	0.1112386	0.0000000	0.0000000	0.1063168	0.0000000
TCGA-CC-A5UD	0.0000000	0.0000000	0.3019476	0.0000000	0.0000000	0.0000000	0.5283109	0.0000000	0.0704792
TCGA-CC-A5UE	0.0000000	0.1903117	0.4258034	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.2648270
TCGA-CC-A7IE	0.0000000	0.0981329	0.4139795	0.0000000	0.0914160	0.0670622	0.1054766	0.0000000	0.1615843
TCGA-CC-A7IF	0.0000000	0.2068446	0.6777681	0.0000000	0.0000000	0.0000000	0.0798178	0.0000000	0.0000000
TCGA-CC-A7IG	0.0000000	0.1453380	0.4282957	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.3550579
TCGA-CC-A7IH	0.0000000	0.1733917	0.2388160	0.0000000	0.2057515	0.0000000	0.3380071	0.0000000	0.0000000
TCGA-CC-A7II	0.0000000	0.1466689	0.0000000	0.0000000	0.0000000	0.1339053	0.0000000	0.0000000	0.5402666

TCGA-CC-A7IJ	0.0000000	0.1870955	0.7539958	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-CC-A7IK	0.0000000	0.1600232	0.2996605	0.0000000	0.1941692	0.0000000	0.3193907	0.0000000	0.0000000
TCGA-CC-A7IL	0.0000000	0.1086681	0.5052843	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.3216847
TCGA-CC-A8HS	0.1116860	0.2235114	0.4118614	0.0000000	0.1296775	0.0000000	0.0000000	0.1004248	0.0000000
TCGA-CC-A8HT	0.0000000	0.2191250	0.4969974	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.1798183
TCGA-CC-A8HU	0.0000000	0.2077722	0.3442225	0.0000000	0.1180082	0.0000000	0.0000000	0.0000000	0.2036769
TCGA-CC-A8HV	0.0000000	0.0000000	0.5532600	0.0000000	0.0000000	0.0000000	0.0675819	0.0000000	0.2051345
TCGA-CC-A9FS	0.0000000	0.0685741	0.5633129	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.2729972
TCGA-CC-A9FU	0.0000000	0.1008945	0.7674467	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0954377
TCGA-CC-A9FV	0.3545002	0.0000000	0.3868419	0.0000000	0.0000000	0.2586579	0.0000000	0.0000000	0.0000000
TCGA-CC-A9FW	0.0000000	0.1014411	0.4421081	0.0000000	0.0609101	0.0935237	0.0939838	0.0666009	0.0851509
TCGA-DD-A113	0.1120880	0.1637431	0.5278610	0.0000000	0.0000000	0.0000000	0.0644808	0.0854468	0.0000000
TCGA-DD-A114	0.0000000	0.1628286	0.3856622	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.4124723
TCGA-DD-A115	0.0000000	0.2583495	0.5509298	0.0000000	0.0000000	0.0000000	0.0000000	0.1263073	0.0644134
TCGA-DD-A116	0.0000000	0.2489123	0.5837223	0.0000000	0.0971395	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-A118	0.1271817	0.0794669	0.7605301	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-A119	0.1107660	0.1142887	0.6660584	0.0000000	0.0777378	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-A11A	0.1332947	0.1152066	0.6637297	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-A11B	0.0000000	0.1200240	0.6424982	0.0000000	0.0965167	0.0000000	0.0000000	0.0883874	0.0000000
TCGA-DD-A11C	0.0862279	0.1183324	0.6844923	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-A11D	0.0000000	0.3427706	0.5005714	0.0000000	0.0810009	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-A1E9	0.0000000	0.0000000	0.8991401	0.0681599	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-A1EB	0.0000000	0.2756917	0.5716605	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-A1EC	0.2379780	0.1135029	0.3332124	0.0000000	0.2262541	0.0000000	0.0000000	0.0890525	0.0000000
TCGA-DD-A1ED	0.0000000	0.2494218	0.3720297	0.0000000	0.1504485	0.0616254	0.0000000	0.0914964	0.0749782

TCGA-DD-A1EE	0.0000000	0.2324733	0.6278237	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0861620
TCGA-DD-A1EF	0.0000000	0.0000000	0.5132742	0.0697471	0.1797279	0.0000000	0.0712471	0.1317183	0.0000000
TCGA-DD-A1EH	0.1903515	0.0000000	0.5547267	0.0988713	0.0000000	0.0000000	0.0000000	0.0000000	0.0752473
TCGA-DD-A1EI	0.0000000	0.2006980	0.4892253	0.0959243	0.1438544	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-A1EJ	0.1043791	0.1300232	0.6901687	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-A1EK	0.0000000	0.1932125	0.6545481	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-A1EL	0.0676923	0.1983642	0.0000000	0.0000000	0.0999375	0.0000000	0.0000000	0.0000000	0.5741528
TCGA-DD-A39V	0.0000000	0.1350044	0.6137798	0.1035336	0.0835460	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-A39W	0.1036295	0.0000000	0.6806649	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.1265985
TCGA-DD-A39X	0.0956561	0.0000000	0.7391071	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0983830
TCGA-DD-A39Z	0.0000000	0.0851456	0.5586464	0.0000000	0.0000000	0.0000000	0.0000000	0.1730688	0.1114782
TCGA-DD-A3A2	0.0000000	0.0720165	0.7308344	0.0821168	0.0000000	0.0000000	0.0000000	0.0000000	0.0792164
TCGA-DD-A3A3	0.1014671	0.0000000	0.7074310	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.1202271
TCGA-DD-A3A4	0.1416722	0.2028298	0.3782575	0.0979293	0.1267834	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-A3A5	0.0000000	0.1225970	0.5953164	0.1529227	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-A3A6	0.0000000	0.1029831	0.4604623	0.2265782	0.1380160	0.0000000	0.0000000	0.0719605	0.0000000
TCGA-DD-A3A7	0.0000000	0.0745508	0.5047170	0.0000000	0.0728305	0.2645990	0.0000000	0.0000000	0.0000000
TCGA-DD-A3A8	0.0000000	0.1498345	0.6611866	0.0000000	0.0897383	0.0613414	0.0000000	0.0000000	0.0000000
TCGA-DD-A3A9	0.0000000	0.0772946	0.6222173	0.0000000	0.1524175	0.0000000	0.0616612	0.0743273	0.0000000
TCGA-DD-A4NA	0.1172943	0.0988465	0.6393897	0.0000000	0.0859673	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-A4NB	0.0000000	0.1302347	0.7912747	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-A4ND	0.0000000	0.0000000	0.8217210	0.0000000	0.0627007	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-A4NE	0.0000000	0.0000000	0.7586409	0.0000000	0.0000000	0.1825629	0.0000000	0.0000000	0.0000000
TCGA-DD-A4NF	0.0000000	0.1918529	0.7649862	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-A4NG	0.0000000	0.0000000	0.8296562	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000

TCGA-DD-A4NH	0.0000000	0.1010783	0.6076765	0.2425589	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-A4NI	0.0000000	0.1875523	0.5842877	0.0000000	0.1353793	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-A4NJ	0.0000000	0.1924637	0.6777156	0.0000000	0.0000000	0.0000000	0.0625425	0.0000000	0.0000000
TCGA-DD-A4NK	0.0000000	0.1050109	0.7391058	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-A4NL	0.1287254	0.0000000	0.6410419	0.0000000	0.0000000	0.1186100	0.0000000	0.0000000	0.0000000
TCGA-DD-A4NN	0.1117670	0.0000000	0.6602561	0.0000000	0.0993047	0.0000000	0.0842824	0.0000000	0.0000000
TCGA-DD-A4NO	0.0890339	0.1239906	0.5551744	0.0000000	0.1497655	0.0000000	0.0746616	0.0000000	0.0000000
TCGA-DD-A4NP	0.1108320	0.0978083	0.4865234	0.0000000	0.1498008	0.0645653	0.0000000	0.0000000	0.0000000
TCGA-DD-A4NQ	0.1276159	0.1168581	0.6478397	0.0000000	0.1076863	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-A4NR	0.0000000	0.1073463	0.7374396	0.0000000	0.0781521	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-A4NS	0.1300991	0.3292342	0.3416227	0.0000000	0.1360571	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-A4NV	0.0681193	0.1052692	0.6663233	0.0000000	0.0710866	0.0000000	0.0892016	0.0000000	0.0000000
TCGA-DD-A73A	0.0000000	0.1857384	0.3357344	0.0820818	0.1245838	0.0000000	0.1459127	0.0000000	0.0717007
TCGA-DD-A73B	0.0633305	0.1137531	0.4440658	0.0756796	0.1265724	0.0877826	0.0000000	0.0000000	0.0000000
TCGA-DD-A73C	0.0000000	0.1060580	0.5046364	0.0000000	0.0000000	0.0956179	0.0673213	0.1315760	0.0000000
TCGA-DD-A73D	0.0000000	0.1190117	0.6773960	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0962323
TCGA-DD-A73E	0.0000000	0.2503995	0.7054033	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-A73F	0.0000000	0.1231660	0.6153145	0.1016061	0.0000000	0.0000000	0.0000000	0.0000000	0.1046229
TCGA-DD-A73G	0.0000000	0.0653106	0.8063492	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-AA3A	0.2039872	0.2604048	0.5160289	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-AAC8	0.0000000	0.0000000	0.1253659	0.0000000	0.0000000	0.0000000	0.7862173	0.0000000	0.0000000
TCGA-DD-AAC9	0.0000000	0.2268769	0.6330452	0.0653567	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-AACB	0.0000000	0.1162781	0.6053610	0.0000000	0.0000000	0.0000000	0.0000000	0.1286777	0.0000000
TCGA-DD-AACC	0.0783005	0.2326659	0.5745677	0.0000000	0.1144659	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-AACD	0.0000000	0.1341619	0.5018320	0.0000000	0.1228237	0.1523679	0.0000000	0.0000000	0.0000000

TCGA-DD-AACE	0.0000000	0.1238390	0.4487248	0.1472830	0.0657449	0.0000000	0.1162521	0.0000000	0.0668540
TCGA-DD-AACF	0.0000000	0.1467575	0.6281662	0.0000000	0.0616383	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-AACG	0.0000000	0.2220922	0.4943165	0.0000000	0.1182581	0.0000000	0.0673027	0.0000000	0.0000000
TCGA-DD-AACH	0.0000000	0.1170617	0.7076381	0.0000000	0.0765754	0.0000000	0.0608192	0.0000000	0.0000000
TCGA-DD-AACI	0.0000000	0.0000000	0.4121190	0.0000000	0.1470793	0.0000000	0.3198363	0.0000000	0.0000000
TCGA-DD-AACJ	0.0876239	0.0000000	0.7150074	0.0000000	0.1494165	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-AACK	0.0000000	0.0000000	0.4186357	0.0000000	0.0000000	0.0000000	0.3596057	0.0000000	0.0779143
TCGA-DD-AACL	0.0000000	0.0000000	0.1354308	0.0000000	0.0000000	0.0000000	0.8234254	0.0000000	0.0000000
TCGA-DD-AACM	0.1775775	0.0765228	0.4970840	0.0000000	0.0983297	0.0000000	0.0630328	0.0874531	0.0000000
TCGA-DD-AACN	0.0934045	0.1157315	0.3810052	0.0000000	0.0889392	0.0664429	0.0893619	0.1246817	0.0000000
TCGA-DD-AACO	0.0000000	0.1028425	0.8734380	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-AACP	0.1381015	0.0000000	0.2859756	0.0000000	0.1056534	0.2492586	0.0666679	0.0691106	0.0787766
TCGA-DD-AACQ	0.0000000	0.0616485	0.2600695	0.0000000	0.0000000	0.0000000	0.6706370	0.0000000	0.0000000
TCGA-DD-AACS	0.0829719	0.0636126	0.6995635	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0969688
TCGA-DD-AACT	0.0000000	0.0681188	0.3615633	0.0000000	0.1167338	0.0000000	0.4053176	0.0000000	0.0000000
TCGA-DD-AACU	0.0000000	0.1222678	0.6049854	0.0000000	0.0000000	0.0000000	0.1853590	0.0000000	0.0000000
TCGA-DD-AACV	0.0609451	0.0000000	0.3920928	0.0728194	0.1596144	0.1829523	0.0000000	0.0000000	0.0818700
TCGA-DD-AACW	0.0000000	0.2661003	0.3422430	0.2278114	0.0000000	0.0773907	0.0000000	0.0000000	0.0000000
TCGA-DD-AACX	0.0000000	0.1626634	0.6395968	0.0000000	0.0948019	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-AACY	0.0000000	0.1222740	0.7566861	0.0000000	0.0000000	0.0713533	0.0000000	0.0000000	0.0000000
TCGA-DD-AACZ	0.0000000	0.0984679	0.2811404	0.1157844	0.0774343	0.0000000	0.3169922	0.0000000	0.0000000
TCGA-DD-AAD0	0.0000000	0.0906109	0.4714250	0.0848470	0.0758719	0.0000000	0.0612248	0.0000000	0.1773221
TCGA-DD-AAD1	0.0774185	0.1751344	0.4439871	0.0000000	0.0852478	0.0000000	0.1048326	0.0000000	0.0000000
TCGA-DD-AAD2	0.0000000	0.1331384	0.6543967	0.0695566	0.0000000	0.0000000	0.0783424	0.0000000	0.0000000
TCGA-DD-AAD3	0.0000000	0.1795761	0.5389799	0.0000000	0.0677814	0.1249609	0.0000000	0.0000000	0.0000000

TCGA-DD-AAD5	0.0609923	0.0676039	0.6374816	0.0000000	0.0000000	0.0619327	0.0714469	0.0000000	0.0000000
TCGA-DD-AAD6	0.0000000	0.0690116	0.5022471	0.0000000	0.0000000	0.2473214	0.0654137	0.0000000	0.0000000
TCGA-DD-AAD8	0.0000000	0.0646575	0.5838676	0.0000000	0.0977455	0.0000000	0.1941274	0.0000000	0.0000000
TCGA-DD-AADA	0.1071261	0.1454601	0.5222335	0.0000000	0.0762819	0.0000000	0.1172314	0.0000000	0.0000000
TCGA-DD-AADB	0.0000000	0.1501845	0.6174893	0.0000000	0.1571103	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-AADC	0.1114452	0.0000000	0.4297334	0.0000000	0.0000000	0.1829573	0.0000000	0.0000000	0.1699350
TCGA-DD-AADD	0.0000000	0.0990074	0.5301752	0.0000000	0.0000000	0.2709131	0.0000000	0.0000000	0.0000000
TCGA-DD-AADE	0.0000000	0.0000000	0.5357996	0.0000000	0.0743430	0.2222072	0.0000000	0.0000000	0.0985124
TCGA-DD-AADF	0.0000000	0.1711545	0.3830110	0.0000000	0.0000000	0.0000000	0.4275705	0.0000000	0.0000000
TCGA-DD-AADG	0.0000000	0.1818334	0.4954324	0.0000000	0.0000000	0.1539589	0.0000000	0.0000000	0.0000000
TCGA-DD-AADI	0.0000000	0.0000000	0.4197561	0.0647872	0.1202662	0.0000000	0.1429650	0.1588513	0.0656022
TCGA-DD-AADJ	0.0000000	0.0000000	0.6189907	0.0000000	0.0000000	0.0000000	0.0729281	0.0000000	0.1637862
TCGA-DD-AADK	0.0000000	0.1602094	0.4597330	0.0000000	0.1960737	0.0922192	0.0000000	0.0000000	0.0000000
TCGA-DD-AADL	0.0000000	0.2003621	0.4476507	0.0000000	0.0703127	0.1653720	0.0000000	0.0000000	0.0000000
TCGA-DD-AADM	0.0000000	0.0000000	0.4126286	0.0000000	0.1095339	0.3055682	0.1150381	0.0000000	0.0000000
TCGA-DD-AADN	0.0620514	0.2177696	0.2895659	0.0000000	0.1304572	0.1409857	0.0000000	0.0966035	0.0625667
TCGA-DD-AADO	0.0000000	0.0000000	0.6810478	0.0000000	0.1084640	0.0000000	0.1022862	0.0000000	0.0640947
TCGA-DD-AADP	0.0941039	0.1184565	0.5187987	0.0000000	0.1428516	0.0985354	0.0000000	0.0000000	0.0000000
TCGA-DD-AADQ	0.0000000	0.1184003	0.7086377	0.0000000	0.0000000	0.0000000	0.0967874	0.0000000	0.0000000
TCGA-DD-AADR	0.0000000	0.1232347	0.7672026	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-AADS	0.0000000	0.1032518	0.4248646	0.0000000	0.0938917	0.0000000	0.3095912	0.0000000	0.0684008
TCGA-DD-AADU	0.0000000	0.0793245	0.3591381	0.0000000	0.1887349	0.0000000	0.1811570	0.0000000	0.1079241
TCGA-DD-AADV	0.0000000	0.0000000	0.4131658	0.0000000	0.0627933	0.2839815	0.0000000	0.0760703	0.0000000
TCGA-DD-AADW	0.0864387	0.2310982	0.2841209	0.0000000	0.1255745	0.0000000	0.1179709	0.1547968	0.0000000
TCGA-DD-AADY	0.0000000	0.0974510	0.7333926	0.1436210	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000

TCGA-DD-AAE0	0.0000000	0.0000000	0.3836015	0.2962143	0.1526505	0.0000000	0.0000000	0.0888034	0.0000000
TCGA-DD-AAE1	0.1343406	0.1020345	0.6158904	0.0000000	0.0910581	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-AAE2	0.0000000	0.1086413	0.7565897	0.0000000	0.0000000	0.0000000	0.1224351	0.0000000	0.0000000
TCGA-DD-AAE3	0.0622338	0.0000000	0.0000000	0.0000000	0.0000000	0.8655785	0.0000000	0.0000000	0.0000000
TCGA-DD-AAE4	0.0931741	0.2611002	0.3611355	0.0000000	0.1164304	0.0000000	0.0652651	0.1028947	0.0000000
TCGA-DD-AAE6	0.0000000	0.0000000	0.8477094	0.0000000	0.0000000	0.0000000	0.0000000	0.1106455	0.0000000
TCGA-DD-AAE8	0.0941212	0.1696861	0.5587267	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-AAE9	0.0000000	0.1405898	0.4840822	0.0000000	0.1633358	0.0000000	0.1430660	0.0000000	0.0000000
TCGA-DD-AAEA	0.0000000	0.0000000	0.4117209	0.0000000	0.0000000	0.4142246	0.0000000	0.0000000	0.0000000
TCGA-DD-AAEB	0.0000000	0.1599951	0.5125352	0.0000000	0.0000000	0.2515457	0.0000000	0.0000000	0.0000000
TCGA-DD-AAED	0.0000000	0.1369097	0.7623904	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-AAEE	0.0783311	0.0000000	0.2743613	0.0000000	0.1374378	0.1829537	0.0000000	0.0000000	0.2012669
TCGA-DD-AAEG	0.1471228	0.1503617	0.4922571	0.0000000	0.0630673	0.1043620	0.0000000	0.0000000	0.0000000
TCGA-DD-AAEH	0.0000000	0.1636741	0.4743376	0.0000000	0.1422621	0.0000000	0.0000000	0.0000000	0.1619957
TCGA-DD-AAEI	0.0000000	0.1091421	0.6087770	0.0000000	0.0000000	0.0000000	0.1595549	0.0000000	0.0000000
TCGA-DD-AAEK	0.0000000	0.0000000	0.5237962	0.0000000	0.0000000	0.2593610	0.0000000	0.0000000	0.1277682
TCGA-DD-AAVP	0.0000000	0.1817431	0.6880059	0.0000000	0.0757449	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-AAVQ	0.0000000	0.0000000	0.7420196	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.1833400
TCGA-DD-AAVR	0.1273520	0.0765986	0.3907813	0.0807374	0.0667573	0.0000000	0.0970337	0.0653083	0.0954314
TCGA-DD-AAVS	0.0794613	0.0706529	0.4274441	0.0000000	0.0000000	0.2188396	0.0000000	0.0000000	0.1668156
TCGA-DD-AAVU	0.0000000	0.0000000	0.7379547	0.0000000	0.0774576	0.0000000	0.1227227	0.0000000	0.0000000
TCGA-DD-AAVV	0.0000000	0.1710665	0.6011710	0.0000000	0.1308278	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-AAVX	0.0000000	0.1588621	0.5153552	0.0609714	0.1117571	0.0000000	0.0613601	0.0916941	0.0000000
TCGA-DD-AAVY	0.0000000	0.1673022	0.5396179	0.0000000	0.1982597	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-DD-AAVZ	0.0726932	0.1107352	0.7974589	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000

TCGA-DD-AAW0	0.0629107	0.0000000	0.5747563	0.0000000	0.0000000	0.0896763	0.0935669	0.0000000	0.0982192
TCGA-DD-AAW1	0.0000000	0.0000000	0.3617666	0.0000000	0.0000000	0.0000000	0.4480977	0.0603438	0.0000000
TCGA-DD-AAW2	0.0000000	0.0000000	0.4650911	0.0000000	0.0780358	0.1176292	0.1360353	0.0000000	0.1021799
TCGA-DD-AAW3	0.0000000	0.0605742	0.6988876	0.0000000	0.0000000	0.0000000	0.0633146	0.0000000	0.0000000
TCGA-ED-A459	0.0000000	0.0642445	0.3181095	0.0000000	0.0000000	0.0000000	0.5853009	0.0000000	0.0000000
TCGA-ED-A4XI	0.0000000	0.0000000	0.4863611	0.0000000	0.1634686	0.2204496	0.0774288	0.0000000	0.0000000
TCGA-ED-A5KG	0.1193062	0.0796809	0.6665549	0.0000000	0.0000000	0.0000000	0.0658698	0.0000000	0.0000000
TCGA-ED-A627	0.0000000	0.0000000	0.4693128	0.0000000	0.4171080	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-ED-A66X	0.2650745	0.0000000	0.6457550	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-ED-A66Y	0.0000000	0.0000000	0.4135728	0.1560852	0.1831549	0.0000000	0.0000000	0.0756182	0.1198495
TCGA-ED-A7PX	0.0998318	0.1820371	0.6252688	0.0928623	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-ED-A7PY	0.1991234	0.0000000	0.3097741	0.0000000	0.1929110	0.0000000	0.0000000	0.0000000	0.2703412
TCGA-ED-A7PZ	0.0000000	0.1418743	0.3352564	0.0000000	0.0000000	0.0000000	0.4310513	0.0000000	0.0000000
TCGA-ED-A7XO	0.0823303	0.2173442	0.2918632	0.1122017	0.1628255	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-ED-A7XP	0.0000000	0.1860359	0.4609797	0.1039279	0.0000000	0.0000000	0.0000000	0.0000000	0.1592481
TCGA-ED-A82E	0.0980743	0.1727244	0.5190487	0.1066357	0.0000000	0.0000000	0.0000000	0.1035170	0.0000000
TCGA-ED-A8O5	0.0000000	0.0687539	0.1845868	0.0898204	0.1563147	0.0000000	0.2474123	0.1496769	0.1034350
TCGA-ED-A8O6	0.0740036	0.1653063	0.2371203	0.2571466	0.1097008	0.0000000	0.0654467	0.0000000	0.0000000
TCGA-ED-A97K	0.1153272	0.0000000	0.0000000	0.3619041	0.0000000	0.2032761	0.1319516	0.0000000	0.1009193
TCGA-EP-A12J	0.1195329	0.2527040	0.5837065	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-EP-A26S	0.0000000	0.1547327	0.7439526	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-EP-A2KA	0.0000000	0.2724827	0.4914474	0.1654710	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-EP-A2KB	0.0000000	0.2643354	0.5432512	0.0000000	0.0000000	0.1295740	0.0000000	0.0000000	0.0000000
TCGA-EP-A2KC	0.0867313	0.2128564	0.5268642	0.0000000	0.0963889	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-EP-A3JL	0.0000000	0.2622342	0.5603283	0.0000000	0.1601411	0.0000000	0.0000000	0.0000000	0.0000000

TCGA-EP-A3RK	0.0000000	0.1076536	0.4376661	0.0825657	0.0000000	0.1031304	0.0787837	0.0985993	0.0000000
TCGA-ES-A2HS	0.0000000	0.1797120	0.5834010	0.1081587	0.0000000	0.0000000	0.0000000	0.0777801	0.0000000
TCGA-ES-A2HT	0.0000000	0.1644378	0.5700975	0.0000000	0.0000000	0.1038910	0.0000000	0.0671065	0.0000000
TCGA-FV-A23B	0.0000000	0.1080687	0.5688231	0.1076095	0.1360578	0.0000000	0.0000000	0.0613300	0.0000000
TCGA-FV-A2QQ	0.0000000	0.2294789	0.5125616	0.0000000	0.0907249	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-FV-A2QR	0.0000000	0.2138099	0.7645903	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-FV-A3I0	0.0784331	0.0000000	0.7358540	0.0000000	0.0000000	0.0000000	0.0000000	0.0800543	0.1056585
TCGA-FV-A3I1	0.0000000	0.2063739	0.6289010	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-FV-A3R2	0.0000000	0.2396154	0.6092234	0.0000000	0.0000000	0.0000000	0.0000000	0.0642980	0.0000000
TCGA-FV-A3R3	0.1326944	0.0000000	0.6037548	0.0000000	0.0636270	0.0000000	0.0000000	0.0000000	0.1283306
TCGA-FV-A495	0.0000000	0.1188981	0.6389614	0.0000000	0.0807232	0.0000000	0.0000000	0.0692499	0.0000000
TCGA-FV-A496	0.0765197	0.1801663	0.6753095	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-FV-A4ZP	0.0829746	0.1568006	0.6280278	0.0000000	0.0000000	0.0000000	0.0000000	0.0945690	0.0000000
TCGA-FV-A4ZQ	0.0000000	0.2072569	0.4232970	0.0000000	0.0000000	0.0879164	0.0000000	0.0000000	0.1654492
TCGA-G3-A25S	0.0000000	0.0000000	0.8700207	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-G3-A25T	0.1201850	0.0738805	0.7151123	0.0000000	0.0908223	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-G3-A25U	0.0000000	0.5248645	0.4110639	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-G3-A25V	0.0000000	0.2327815	0.3888535	0.0000000	0.1086052	0.1679655	0.0000000	0.0000000	0.0000000
TCGA-G3-A25W	0.0000000	0.1621090	0.5303178	0.0000000	0.0967110	0.1391912	0.0000000	0.0000000	0.0000000
TCGA-G3-A25Y	0.0000000	0.3005589	0.5541184	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0951279
TCGA-G3-A25Z	0.0767530	0.4028471	0.2374279	0.0000000	0.1425343	0.0000000	0.0000000	0.0758803	0.0000000
TCGA-G3-A3CG	0.0687687	0.1580002	0.5206956	0.0000000	0.0000000	0.0738913	0.0929545	0.0000000	0.0000000
TCGA-G3-A3CH	0.0673779	0.0000000	0.5262845	0.1384595	0.0000000	0.0000000	0.0969479	0.0781423	0.0927879
TCGA-G3-A3CI	0.0000000	0.2518694	0.4171757	0.0734322	0.2015418	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-G3-A3CK	0.0000000	0.0000000	0.7028010	0.0000000	0.1789509	0.0000000	0.0000000	0.0000000	0.0668325

TCGA-G3-A5SI	0.0710400	0.0974476	0.5481888	0.0000000	0.1338635	0.0000000	0.0000000	0.0000000	0.1100315
TCGA-G3-A5SJ	0.0000000	0.1341939	0.6254377	0.0000000	0.0878709	0.1360318	0.0000000	0.0000000	0.0000000
TCGA-G3-A5SK	0.0000000	0.2134293	0.7101938	0.0000000	0.0763770	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-G3-A5SL	0.0000000	0.1042836	0.6969400	0.0000000	0.0000000	0.0769095	0.0000000	0.0000000	0.0000000
TCGA-G3-A5SM	0.0000000	0.1389537	0.6494341	0.1508819	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-G3-A6UC	0.0000000	0.1908082	0.6045166	0.0000000	0.0695775	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-G3-A7M5	0.0000000	0.0727242	0.3503243	0.0000000	0.0000000	0.0000000	0.4527639	0.0000000	0.0000000
TCGA-G3-A7M6	0.0000000	0.0720166	0.8656397	0.0000000	0.0623437	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-G3-A7M7	0.0000000	0.1424366	0.7937578	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-G3-A7M8	0.0000000	0.0000000	0.2893529	0.2157398	0.1856922	0.1145985	0.0775935	0.0000000	0.0701481
TCGA-G3-A7M9	0.0000000	0.2851278	0.4950697	0.0000000	0.0614331	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-G3-AAUZ	0.0817630	0.2532658	0.3164905	0.0000000	0.2038178	0.0000000	0.0000000	0.0000000	0.0897033
TCGA-G3-AAV0	0.0000000	0.0878398	0.6811910	0.0000000	0.0000000	0.0000000	0.1030539	0.0662917	0.0000000
TCGA-G3-AAV1	0.0000000	0.0607808	0.6520718	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.1381642
TCGA-G3-AAV2	0.0000000	0.0000000	0.4996423	0.1727968	0.0802590	0.0000000	0.1175394	0.0000000	0.1165411
TCGA-G3-AAV3	0.0000000	0.1767964	0.5587914	0.1035147	0.0000000	0.0000000	0.0998180	0.0000000	0.0000000
TCGA-G3-AAV4	0.0000000	0.2320077	0.3506878	0.0998740	0.0000000	0.1628334	0.0000000	0.1421205	0.0000000
TCGA-G3-AAV5	0.0000000	0.0950545	0.5937938	0.1236823	0.0000000	0.0000000	0.0000000	0.0000000	0.1199012
TCGA-G3-AAV6	0.0000000	0.2191245	0.6398106	0.1006720	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-G3-AAV7	0.0000000	0.1417963	0.6256412	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0933749
TCGA-GJ-A3OU	0.0620810	0.0709595	0.6135269	0.0000000	0.0675968	0.0000000	0.0000000	0.0000000	0.0841214
TCGA-GJ-A6C0	0.0000000	0.0000000	0.5058840	0.0000000	0.0000000	0.0000000	0.0000000	0.2779284	0.1254045
TCGA-GJ-A9DB	0.0000000	0.0656052	0.7032975	0.0000000	0.0000000	0.0000000	0.0000000	0.0603430	0.0914345
TCGA-HP-A5MZ	0.0000000	0.0932536	0.8156476	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-HP-A5N0	0.0000000	0.2474589	0.5081548	0.0000000	0.1333040	0.0000000	0.0000000	0.0000000	0.0000000

TCGA-K7-A5RF	0.0918083	0.0000000	0.5232341	0.0000000	0.1304168	0.1237496	0.0000000	0.0000000	0.0767534
TCGA-K7-A5RG	0.0705774	0.0989464	0.4962278	0.0000000	0.1722469	0.0000000	0.1449706	0.0000000	0.0000000
TCGA-K7-A6G5	0.0783779	0.2211517	0.4591710	0.0000000	0.0000000	0.1480653	0.0000000	0.0000000	0.0000000
TCGA-K7-AAU7	0.0610477	0.0000000	0.8303845	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-KR-A7K0	0.0000000	0.2475584	0.4489512	0.0000000	0.1040273	0.0000000	0.0000000	0.0714902	0.0000000
TCGA-KR-A7K2	0.0000000	0.1733205	0.5007188	0.0699626	0.1632223	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-KR-A7K7	0.1626467	0.0896455	0.6957054	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-KR-A7K8	0.1137487	0.2159784	0.4745594	0.0000000	0.1651538	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-LG-A6GG	0.0000000	0.1445111	0.4217054	0.0000000	0.0000000	0.0000000	0.3598508	0.0000000	0.0000000
TCGA-LG-A9QC	0.0000000	0.0980585	0.4572404	0.0000000	0.0000000	0.1552642	0.0000000	0.0722487	0.0780328
TCGA-LG-A9QD	0.0000000	0.1440802	0.6077592	0.0000000	0.0889251	0.0000000	0.0000000	0.0787114	0.0000000
TCGA-MI-A75C	0.0000000	0.1626794	0.5886810	0.0000000	0.0000000	0.1024261	0.0000000	0.0000000	0.0000000
TCGA-MI-A75E	0.0000000	0.0797412	0.5860902	0.0000000	0.1180038	0.0703146	0.0925328	0.0000000	0.0000000
TCGA-MI-A75G	0.0000000	0.2268281	0.3754222	0.0000000	0.0660047	0.1637299	0.0000000	0.0000000	0.0688868
TCGA-MI-A75H	0.0000000	0.2121219	0.6667718	0.0000000	0.0951415	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-MI-A75I	0.0000000	0.1252891	0.6287153	0.0000000	0.0679188	0.0000000	0.1325713	0.0000000	0.0000000
TCGA-MR-A520	0.0000000	0.0000000	0.6109167	0.0690119	0.0000000	0.1990694	0.0000000	0.0000000	0.0980507
TCGA-MR-A8JO	0.0000000	0.0000000	0.8450571	0.0000000	0.0000000	0.0732677	0.0000000	0.0000000	0.0000000
TCGA-NI-A4U2	0.0000000	0.0807200	0.7660890	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-NI-A8LF	0.0000000	0.1723343	0.7356102	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-O8-A75V	0.0000000	0.1970186	0.5772375	0.0000000	0.0000000	0.1197939	0.0000000	0.0000000	0.0000000
TCGA-PD-A5DF	0.0933469	0.0697730	0.7363164	0.0000000	0.0000000	0.0000000	0.0673877	0.0000000	0.0000000
TCGA-QA-A7B7	0.0000000	0.2174420	0.5546145	0.0000000	0.1065912	0.0000000	0.0000000	0.0000000	0.0693099
TCGA-RC-A6M3	0.0775263	0.0000000	0.6510599	0.0000000	0.1089683	0.0731798	0.0000000	0.0000000	0.0661456
TCGA-RC-A6M4	0.0000000	0.2136037	0.2806473	0.0000000	0.1314808	0.0000000	0.2869433	0.0000000	0.0000000

TCGA-RC-A6M5	0.2902246	0.0850658	0.3889362	0.0000000	0.1746608	0.0611126	0.0000000	0.0000000	0.0000000
TCGA-RC-A6M6	0.0000000	0.2540797	0.5342506	0.0000000	0.0000000	0.0000000	0.0735581	0.0000000	0.0969460
TCGA-RC-A7S9	0.0861715	0.1897423	0.3517980	0.0000000	0.1467783	0.0786486	0.0000000	0.0812081	0.0656531
TCGA-RC-A7SB	0.0000000	0.0000000	0.4522133	0.0738785	0.1364588	0.0991952	0.1140527	0.0000000	0.1075717
TCGA-RC-A7SF	0.0000000	0.0916563	0.3447191	0.1201090	0.1065505	0.0960658	0.1090050	0.0000000	0.1146616
TCGA-RC-A7SH	0.1276970	0.0853933	0.5243718	0.0000000	0.1376324	0.0000000	0.0790028	0.0000000	0.0000000
TCGA-RC-A7SK	0.0000000	0.1490753	0.3560686	0.0000000	0.1354675	0.2694106	0.0899779	0.0000000	0.0000000
TCGA-RG-A7D4	0.0000000	0.1284842	0.7852263	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-T1-A6J8	0.0000000	0.2763034	0.4039411	0.0913807	0.0000000	0.1074462	0.0000000	0.0000000	0.1126891
TCGA-UB-A7MA	0.1457204	0.0000000	0.8051775	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-UB-A7MB	0.0000000	0.1502488	0.1196164	0.0000000	0.2637019	0.0000000	0.4664330	0.0000000	0.0000000
TCGA-UB-A7MC	0.0000000	0.0923638	0.8728328	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-UB-A7MD	0.0963368	0.1837034	0.6122623	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-UB-A7ME	0.0866608	0.1134740	0.4485529	0.0000000	0.0000000	0.0822497	0.2690626	0.0000000	0.0000000
TCGA-UB-A7MF	0.1001341	0.1496332	0.6808627	0.0000000	0.0600870	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-UB-AA0U	0.0804736	0.1099734	0.6581891	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-UB-AA0V	0.1664229	0.2455485	0.2571290	0.0000000	0.1388436	0.0000000	0.0000000	0.0899837	0.1020723
TCGA-WJ-A86L	0.0000000	0.1872276	0.6736652	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0807599
TCGA-WQ-A9G7	0.1066959	0.1231928	0.1018698	0.4301853	0.2380562	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-WQ-AB4B	0.0667395	0.0716655	0.8046608	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-WX-AA44	0.0689667	0.1499158	0.6015723	0.0000000	0.1193749	0.0000000	0.0601702	0.0000000	0.0000000
TCGA-WX-AA46	0.0000000	0.1417403	0.6747799	0.0000000	0.1181987	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-WX-AA47	0.0000000	0.0000000	0.8580247	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-XR-A8TC	0.1164578	0.2476712	0.4986745	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-XR-A8TD	0.0000000	0.2112267	0.7367336	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000

TCGA-XR-A8TE	0.3303607	0.3353755	0.0000000	0.0000000	0.3013669	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-XR-A8TF	0.0000000	0.0000000	0.4094200	0.0000000	0.0000000	0.0000000	0.2972022	0.0000000	0.1768365
TCGA-XR-A8TG	0.0000000	0.2167477	0.6527436	0.0000000	0.0947337	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-YA-A8S7	0.0997945	0.0000000	0.6137209	0.0000000	0.1501757	0.0000000	0.0000000	0.0000000	0.0775668
TCGA-ZP-A9CV	0.0000000	0.0000000	0.7389857	0.0000000	0.1132547	0.0000000	0.0000000	0.0000000	0.0691245
TCGA-ZP-A9CY	0.0000000	0.0961957	0.4898542	0.1136893	0.1643652	0.0000000	0.0000000	0.0000000	0.1358956
TCGA-ZP-A9CZ	0.1244145	0.0000000	0.4460326	0.0662259	0.1392496	0.0000000	0.0655858	0.0000000	0.0945970
TCGA-ZP-A9D0	0.0674545	0.0000000	0.7175249	0.0000000	0.0000000	0.0000000	0.0888476	0.0000000	0.0755883
TCGA-ZP-A9D1	0.0000000	0.1420500	0.3312420	0.0000000	0.2747647	0.0000000	0.2028934	0.0000000	0.0000000
TCGA-ZP-A9D2	0.0733056	0.1235485	0.4495570	0.0000000	0.2657461	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-ZP-A9D4	0.0845841	0.0000000	0.5072131	0.0000000	0.1133006	0.0000000	0.0000000	0.0000000	0.1900242
TCGA-ZS-A9CD	0.0000000	0.1603207	0.6038868	0.0000000	0.0000000	0.0000000	0.0000000	0.1065871	0.0000000
TCGA-ZS-A9CE	0.0805492	0.1398992	0.7209159	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
TCGA-ZS-A9CF	0.0000000	0.1626211	0.6408405	0.0000000	0.0000000	0.0000000	0.0948908	0.0000000	0.0000000
TCGA-ZS-A9CG	0.0000000	0.0865485	0.7076086	0.0000000	0.0000000	0.0000000	0.1081271	0.0800158	0.0000000

Table S9: Shared and private copy number peaks across Asian and European cohorts

Cytoband	Type	Cohort	q.values	Gene.Symbol	Status
1q21.1	amp	ASIAN	0.0000811	NA	Shared peak
1q22	amp	ASIAN	0.0024811	NA	Shared peak
2p11.1	amp	ASIAN	0.0000000	NA	Shared peak
3q29	amp	ASIAN	0.1480100	BDH1	Shared peak
4p14	amp	ASIAN	0.0005004	NA	Private amp
5p15.33	amp	ASIAN	0.0000026	TERT	Shared peak
5q35.2	amp	ASIAN	0.0885590	NSD1	Shared peak
6p25.3	amp	ASIAN	0.0688830	NA	Private amp
6p21.32	amp	ASIAN	0.1870800	NA	Private amp
6q12	amp	ASIAN	0.0815820	EYS	Shared peak
6q12	amp	ASIAN	0.0815820	PHF3	Shared peak
7q36.3	amp	ASIAN	0.1308200	PTPRN2	Shared peak
8q24.13	amp	ASIAN	0.0163650	NA	Private amp
11q13.3	amp	ASIAN	0.0000000	CCND1	Shared peak
11q13.3	amp	ASIAN	0.0000000	FGF19	Shared peak
13q34	amp	ASIAN	0.0000000	ING1	Shared peak
14q11.2	amp	ASIAN	0.0294510	CHD8	Private amp
14q11.2	amp	ASIAN	0.0294510	PRMT5	Private amp
16q11.2	amp	ASIAN	0.0000000	NA	Shared peak
17p11.2	amp	ASIAN	0.0158860	NA	Private amp
17q25.3	amp	ASIAN	0.0521120	CBX8	Private amp
17q25.3	amp	ASIAN	0.0521120	FOXK2	Private amp
18q23	amp	ASIAN	0.1223200	NA	Private amp
19p13.12	amp	ASIAN	0.0454010	NOTCH3	Private amp
19p13.12	amp	ASIAN	0.0454010	BRD4	Private amp
19p13.11	amp	ASIAN	0.0719940	NA	Private amp
19q13.2	amp	ASIAN	0.0499560	PAF1	Private amp
1p36.31	del	ASIAN	0.0000000	RPL22	Shared peak
2q31.2	del	ASIAN	0.2008900	NFE2L2	Shared peak
2q37.3	del	ASIAN	0.0077107	ING5	Shared peak
4q13.3	del	ASIAN	0.0171610	COX18	Shared peak
4q13.3	del	ASIAN	0.0171610	ALB	Shared peak
4q35.1	del	ASIAN	0.0000000	NA	Shared peak

6q27	del	ASIAN	0.0000024	PHF10	Shared peak
8p23.1	del	ASIAN	0.0171610	NA	Shared peak
9p21.3	del	ASIAN	0.0003187	CDKN2A	Shared peak
9p21.3	del	ASIAN	0.0003187	CDKN2B	Shared peak
9p11.2	del	ASIAN	0.1365400	NA	Shared peak
11q13.1	del	ASIAN	0.2174100	NA	Private del
11q22.3	del	ASIAN	0.1893600	ATM	Shared peak
12q24.32	del	ASIAN	0.0612340	NA	Shared peak
13q12.11	del	ASIAN	0.0248170	MPHOSPH8	Private del
13q12.11	del	ASIAN	0.0248170	XPO4	Private del
13q14.2	del	ASIAN	0.0000000	RB1	Shared peak
14q32.33	del	ASIAN	0.0023367	TDRD9	Shared peak
14q32.33	del	ASIAN	0.0023367	AKT1	Shared peak
15q11.2	del	ASIAN	0.0000489	NA	Private del
15q26.3	del	ASIAN	0.0116760	IGF1R	Shared peak
16p13.3	del	ASIAN	0.0054228	AXIN1	Shared peak
16p13.3	del	ASIAN	0.0054228	TSC2	Shared peak
16p13.3	del	ASIAN	0.0054228	RBFOX1	Shared peak
16q11.2	del	ASIAN	0.0000000	NA	Shared peak
16q24.3	del	ASIAN	0.0102880	NA	Private del
22q11.1	del	ASIAN	0.0916450	NA	Shared peak
1p36.13	amp	EUROPEAN	0.0178140	NA	Private amp
1q21.1	amp	EUROPEAN	0.0000032	NA	Shared peak
1q21.3	amp	EUROPEAN	0.0000000	MCL1	Shared peak
1q21.3	amp	EUROPEAN	0.0000000	ARNT	Shared peak
1q21.3	amp	EUROPEAN	0.0000000	KCNN3	Shared peak
1q44	amp	EUROPEAN	0.0004106	SMYD3	Private amp
2p24.2	amp	EUROPEAN	0.0567610	NA	Private amp
2p11.1	amp	EUROPEAN	0.0000320	NA	Shared peak
3q26.31	amp	EUROPEAN	0.1040200	NA	Shared peak
4q13.2	amp	EUROPEAN	0.1569700	NA	Private amp
5p15.33	amp	EUROPEAN	0.0043469	TERT	Shared peak
5q35.3	amp	EUROPEAN	0.1079300	PRELID1	Shared peak
6p21.32	amp	EUROPEAN	0.0041401	NA	Private amp
6p21.1	amp	EUROPEAN	0.0004854	CRIP3	Private amp
6p11.1	amp	EUROPEAN	0.0000946	NA	Shared peak

7q31.2	amp	EUROPEAN	0.0010993	MET	Private amp
7q36.3	amp	EUROPEAN	0.1250200	PTPRN2	Shared peak
8q21.2	amp	EUROPEAN	0.0708710	NA	Private amp
8q24.3	amp	EUROPEAN	0.0333420	PARP10	Private amp
8q24.3	amp	EUROPEAN	0.0333420	CYC1	Private amp
9q34.3	amp	EUROPEAN	0.2307700	NOTCH1	Private amp
11q13.3	amp	EUROPEAN	0.0000000	CCND1	Shared peak
11q13.3	amp	EUROPEAN	0.0000000	FGF19	Shared peak
12p13.33	amp	EUROPEAN	0.2231600	KDM5A	Private amp
12q14.1	amp	EUROPEAN	0.0567610	CDK4	Private amp
13q34	amp	EUROPEAN	0.0000185	ING1	Shared peak
15q26.3	amp	EUROPEAN	0.0243830	IGF1R	Private amp
16q11.2	amp	EUROPEAN	0.0000000	NA	Shared peak
17p11.2	amp	EUROPEAN	0.0708710	NA	Private amp
17q25.1	amp	EUROPEAN	0.0001172	NA	Private amp
19p13.2	amp	EUROPEAN	0.0101320	MRPL4	Private amp
19p13.2	amp	EUROPEAN	0.0101320	KEAP1	Private amp
20q13.2	amp	EUROPEAN	0.0579580	ZNF217	Private amp
20q13.2	amp	EUROPEAN	0.0579580	ZNF217	Private amp
22q13.1	amp	EUROPEAN	0.0995880	TNRC6B	Private amp
1p36.31	del	EUROPEAN	0.0000000	RPL22	Shared peak
1p36.13	del	EUROPEAN	0.0000000	NA	Private del
2q22.1	del	EUROPEAN	0.0003500	LRP1B	Shared peak
2q37.3	del	EUROPEAN	0.0105250	ING5	Shared peak
3p21.1	del	EUROPEAN	0.0017145	BAP1	Private del
3p21.1	del	EUROPEAN	0.0017145	PBRM1	Private del
3p13	del	EUROPEAN	0.0093281	NA	Private del
3q29	del	EUROPEAN	0.0178180	BDH1	Private del
4p16.3	del	EUROPEAN	0.0450180	FGFR3	Private del
4p16.3	del	EUROPEAN	0.0450180	LETM1	Private del
4q21.23	del	EUROPEAN	0.0000010	NA	Shared peak
4q22.1	del	EUROPEAN	0.0000001	FAM190A	Shared peak
4q23	del	EUROPEAN	0.0000510	ADH1B	Shared peak
4q35.2	del	EUROPEAN	0.0000049	FAT1	Shared peak
4q35.2	del	EUROPEAN	0.0000049	FRG1	Shared peak
5q14.3	del	EUROPEAN	0.1750000	NA	Private del

6q16.1	del	EUROPEAN	0.0382510	NA	Private del
6q27	del	EUROPEAN	0.0051302	PHF10	Shared peak
7p22.3	del	EUROPEAN	0.1886400	NA	Private del
7q36.3	del	EUROPEAN	0.1268900	PTPRN2	Private del
8p23.1	del	EUROPEAN	0.0053602	NA	Shared peak
9p24.2	del	EUROPEAN	0.2483400	NA	Private del
9p21.2	del	EUROPEAN	0.0265140	NA	Shared peak
9p12	del	EUROPEAN	0.0018691	NA	Shared peak
9q34.3	del	EUROPEAN	0.0137220	NOTCH1	Private del
10q26.11	del	EUROPEAN	0.0002388	NA	Private del
11p15.5	del	EUROPEAN	0.0004843	SIRT3	Shared peak
11p15.5	del	EUROPEAN	0.0004843	PHRF1	Shared peak
11q25	del	EUROPEAN	0.1797900	NA	Shared peak
12p13.31	del	EUROPEAN	0.1330100	NA	Private del
12q24.33	del	EUROPEAN	0.0308450	NA	Shared peak
13q14.2	del	EUROPEAN	0.0000000	RB1	Shared peak
13q22.2	del	EUROPEAN	0.0000017	NA	Private del
14q32.33	del	EUROPEAN	0.0000000	TDRD9	Shared peak
14q32.33	del	EUROPEAN	0.0000000	AKT1	Shared peak
15q26.1	del	EUROPEAN	0.1801300	NA	Shared peak
16p13.3	del	EUROPEAN	0.0003068	AXIN1	Shared peak
16p13.3	del	EUROPEAN	0.0003068	TSC2	Shared peak
16p13.3	del	EUROPEAN	0.0003068	RBFOX1	Shared peak
16q11.2	del	EUROPEAN	0.0000000	NA	Shared peak
17p11.2	del	EUROPEAN	0.0759250	NA	Private del
19p13.3	del	EUROPEAN	0.0341260	STK11	Private del
21q21.3	del	EUROPEAN	0.0662600	NA	Private del
21q22.3	del	EUROPEAN	0.2170700	PTTG1IP	Private del
22q11.1	del	EUROPEAN	0.0343870	NA	Shared peak
22q13.33	del	EUROPEAN	0.0009447	BRD1	Private del
22q13.33	del	EUROPEAN	0.0009447	HDAC10	Private del