DEVELOPING METHODS TO DETECT DNA REPAIR DEFECTS IN A DIVERSE POPULATION

Markia A. Smith

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Pathology and Laboratory Medicine (Pathobiology and Translational Science) in the School of Medicine.

Chapel Hill
2022

Approved by:

Russell Broaddus

Melissa A. Troester

Katherine A. Hoadley

Michael Love

Ugwuji Maduekwe

ABSTRACT

Markia A. Smith: Developing methods to detect DNA repair defects in a diverse population (Under the direction of Melissa A. Troester and Katherine A. Hoadley)

Primary liver cancer, including hepatocellular carcinoma (HCC) and cholangiocarcinoma (CCA), is one of the fastest growing cancer types. Although HCCs and CCAs are anatomically co-localized, they have distinct etiologic and genomic characteristics that varies vastly in clinical outcome and response to therapy. The Cancer Genome Atlas (TCGA) identified a subset of HCC tumors with CCA genomic features, suggesting these tumors may be a separate HCC class based on their relatedness to CCA. While multiple groups have performed molecular characterization of liver tumors in an effort to identify subtypes, few have investigated beyond gene expression and/or mutations, in an integrated HCC and CCA analysis, or in association with outcomes and liver specific processes (e.g., liver regeneration). This is particularly important because DNA repair dysfunction and liver regeneration are tightly coupled processes implicated in impaired genomic integrity and hepatocarcinogenesis. Particularly, dysregulation of these pathways may be linked to chemoresistance. Given the lack of targeted therapeutic modalities for HCC and ongoing efforts to reduce recurrence, further characterizing and subdividing HCC based on multiple pathway interactions and identification of biomarkers that are associated with repair-mediated survival represents an unmet clinical need.

To address this knowledge gap, the current body of work leveraged two TCGA studies, HCC and CCA, both detailed in molecular, histological, and clinical data across multiple platforms. Through a multi-omic approach, chapter 2 of this study characterized distinct HCC subclasses utilizing an integrated TCGA HCC and CCA dataset to gain insights into biology. Chapter 3

leveraged RNA expression profiling to investigate DNA repair in association with mitotic and regenerative signatures, and clinicopathologic variables in TCGA HCC study.

This work identified three molecularly distinct HCC subclasses associated with viral infection and progression-free survival. In addition, RNA-based classification of DNA repair identified heterogeneity of repair pathways in HCC tumors, and a subset of tumors with substantial disrupted liver biology and poor outcomes. Collectively, this work contributes novel findings about HCC features and repair dysfunction that dictate prognosis, and highlights the importance of developing class specific biomarkers and targeted therapies.

In memory of my grandmother, I could not have done this without you. You always supported my love of science. To all the Black disabled queers, we deserve to be here.

ACKNOWLEDGEMENTS

I would like to thank my mentors, Katherine Hoadley and Melissa Troester, for providing an inclusive welcoming environment, for nurturing my scientific curiosity through innovative and meaningful projects, and for their continued support and guidance. I appreciate their encouragement and patience throughout my journey, and this will stay with me as I continue my career. I would also like to thank the Hoadley and Troester Labs, both current and former, for their scientific contributions to these projects and personal support that helped me grow as a scientist and made the experience enjoyable.

I am extremely appreciative of my committee members, Ugwuji Maduekwe, Russell Broaddus, and Michael Love, for their support and critical analysis over this journey.

A special thank you to the Black science community for their camaraderie, the gripe sessions, accountability groups, reassurance, and cheerleading during this hard but rewarding journey. BlackInGenetics started out as a social media event, and now, has turned into an amazing community of scientific and personal support.

Finally, I am thankful for the unending support and love from my chosen family and friends. To my amazing friends, I wish I had the space to name you all here. Thank you for the phone calls, memes, shared meals, celebrations, and trips. To my ride or die Aliyah, you've been here through everything with me and continue to stay down for me. To Bret, thanks for being my own personal cheerleader and staying up past your bedtime to motivate me. I will always unreservedly support all of your endeavors. To my sister Pern, you've been the constant source of enthusiasm and laughter. I'm glad you'll always bring the tambourines. Thank you for always believing in me during whatever challenge. I love you all.

PREFACE

Chapter 1 provides an introductory overview of DNA repair mechanisms and the role of DNA damage in cancer. I also provide a primer of hepatocellular carcinoma (HCC), including epidemiology, risk factors, pathophysiology, genetics and molecular signatures, and treatment. Any figures included in this section were created for the purpose of this document.

The research presented in Chapter 2 is a published manuscript in Nature Communications Biology that characterized rare genomic subclasses of HCC, integrating across multiple datatypes – transcriptomics, mutations, copy number – to reveal underlying etiology. My contributions to the study include experimental design, performed experiments, data analysis and writing. Co-authors include Jeffrey S. Damrauer, Vonn Walter, Aatish Thennavan, Lisle E. Mose, Sara R. Selitsky, and Katherine A. Hoadley. Jeffrey Damrauer and Katherine Hoadley conceived and designed the study. Katherine Hoadley provided guidance and oversight throughout the study. Jeffrey Damrauer and Katherine Hoadley designed the experiments and contributed to writing the paper. Jeffrey Damrauer, Vonn Walter, Aatish Thennavan, Lisle Mose, Sara R. Selitsky performed experiments and analyzed data. All authors reviewed and edited the paper. Note that the text included here may slightly deviate from the final published version.

The study presented in Chapter 3 is a manuscript in preparation that investigates DNA repair defects in The Cancer Genome Atlas (TCGA) HCC study to understand DNA repair in the context of liver homeostasis and how these defects may play a role in prognosis. My contributions to this work include experimental design, data analysis and writing of the

manuscript. Co-authors include Sarah C. Van Alsten, Andrea Walens, Jeffrey S. Damrauer, Ugwuji Maduekwe, Russell Broaddus, Michael I. Love, Melissa A. Troester, and Katherine A. Hoadley. Sarah Van Alsten and Andrea Walens contributed expertise in developing classifiers. Jeffrey Damrauer provided intellectual input and aid with accessing of publicly available data. Ugwuji Maduekwe and Russell Broaddus provided clinical expertise in HCC pathogenesis and treatment management. Michael Love contributed expertise in statistical analyses. Melissa Troester and Katherine Hoadley provided supervision, guidance, and scholarly input throughout the study.

Chapter 4 provides an integrative summary of studies presented in Chapters 2 and 3. This chapter also outlines future directions for studies to further characterize hepatocellular carcinoma pathway interactions and the potential translational implications. Lastly, to reduce redundancy, a single reference list can be found at the end of this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Alt-EJ	Alternative End Joining
BER	Base excision repair
CCA	Cholangiocarcinoma
CI	Confidence Interval
ClaNC	Classification to the nearest centroid
CLCHCC	Cholangio-like group of tumors
CNA	Copy number alteration
CNV	Copy number variation
COSMIC	Catalogue Of Somatic Mutations In Cancer
CS	Cosine similarity
CTA	Cancer Testis Antigens
CYP	Cytochrome P450
DDR	DNA damage response
EMT	Epithelial-Mesenchymal Transition
ES	Embryonic stem
FA	Fanconi Anemia
FC	Fold change
FDR	False discovery rate

GDC	Genomic Data Commons
GISTIC	Genomic Identification of Significant Targets in Cancer
GSEA	Gene Set Enrichment
HBV	hepatitis B virus
HCC	Hepatocellular carcinoma
HCV	hepatitis C virus
HR	Hazard Ratio
HRD	Homologous recombination deficiency
ICB	Immune checkpoint blockade
ICI	Immune checkpoint inhibitor
IQR	Interquartile range
LOH	Loss of Heterozygosity
LST	Large scale transitions
MMR	Mismatch repair
Mut	Mutations
NAFLD	Nonalcoholic fatty liver disease
NASH	Non-alcoholic steatohepatitis
NBD	Normal bile duct
NC	North Carolina

NER	Nucleotide Excision Repair
NES	Normalized Enrichment Scores
NHEJ	Nonhomologous end joining
NtAI	Telomeric allelic imbalance score
OS	Overall survival
PARP	Poly ADP ribose polymerase
PFS	Progression-free survival
RFD	Relative Frequency Difference
SBS	Single base substitution
SD	Standard deviation
SNV	Single nucleotide variants
TCGA	The Cancer Genome Atlas
TKI	Tyrosine kinase inhibitor
TLS	Translesion Synthesis
TME	Tumor microenvironment
US	United States
VAF	Variant allele frequency
WT	Wildtype

CHAPTER 1: INTRODUCTION

1.1 Hepatocellular carcinoma

Liver cancer remains a global health issue with an estimated incidence of >1 million cases by 2025[1,2]. Hepatocellular carcinoma (HCC) is the most common type of liver cancer, comprising ~90% of cases. Over 90% of HCC cases occur in the setting of chronic liver disease. Despite advances in the management of HCC, prognosis remains poor with a 5-year survival rate of 18%, largely due to higher rates of diagnosis at advanced stages[3,4].

1.1.1 Epidemiology

The incidence of HCC in the US has tripled over the past four decades, from 14 million in 2012 and is expected to rise to 22 million in the next two decades[5]. Hepatocellular is the 5th most common cancer in males and the 7th in females, with 41,260 new cases diagnosed each year[6]. 80% of HCC cases occur in sub-Saharan Africa and East Asia, and the major risk factors are hepatitis B virus (HBV) infection and aflatoxin exposure[7]. Whereas, in the United States (US), Europe, and Japan the increasing occurrence of hepatitis C virus (HCV) infection is the primary risk factor. Additionally, in the US, the increased prevalence of obesity, nonalcoholic fatty liver disease (NAFLD), nonalcoholic steatohepatitis (NASH), type 2 diabetes, and the other elements of metabolic syndrome contribute to HCC incidence[7].

HCC incidence in the US continues to rise with a disproportionate impact on racial and ethnic marginalized populations. While there have been decreasing rates of HCC seen among Asian/Pacific Islanders, the prevalence in Black people, Hispanics, and individuals older than 65 years remains high and continues to rise[8,9]. HCC incidence progressively increases as individuals reach above the age of 65, with the median age at diagnosis in the US between 60

to 65 years of age[6]. Incidence rates among males two to four-fold higher than rates among females[10]. In the US, the age-adjusted incidence rate among males was 14.4 per 100,000 while the rates among females was 5.2 per 100,000[6]. This sex disparity is more than likely due to the complex interaction from differences in behavioral risk factors, metabolic factors, tumor biology, and therapeutic modalities. Further characterization of mechanisms across biological, behavioral, socioenvironmental, and healthcare systems are necessary to reduce disparities in risk and improve outcomes.

1.1.2 Risk factors

As previously mentioned, several risk factors have been implicated in the development and progression of HCC. The major risk factors for HCC include Hepatitis B virus (HBV) and Hepatitis C virus (HCV) infection, cirrhosis, alcohol abuse, fatty liver disease, diabetes, obesity-related non-alcoholic steatohepatitis, and exposure to aflatoxin. During the recent years, metabolic syndrome has become more frequent and a major cause of HCC in the US. Additionally, mutational signatures analyses have proven aristolochic acid and tobacco as pathogenic cofactors in HCC[11,12]. This dissertation will focus on characterizing HCC based on multiple pathway interactions, tumor characteristics and risk factors, as they may be reflective of the molecular and cellular consequences of pathway dysfunction and genetic instability that predict prognosis.

1.1.2.1 HBV infection

HBV viral load is a major contributor to the incidence of poor prognosis and existing disparities by race/ethnicity and sex. Over 350 million people across the world are infected with or have been infected with HBV, among which 75% of the infected people are Asian[13]. Liver tumors in Asians exhibit higher tumor grade, presence of macrovascular invasion, and the lowest fraction of differentiated samples[11,14]. Approximately, 55% of HCCs are caused by HBV infection, with rates of incidence between 340 to 804 per 100,000 HBV positive males per year

and 120 to 178 per 100,000 HBV positive females[13,15]. HBV-related HCC is a complex, stepwise process that progresses over time. HBV interacts with endogenous mutagens, the inflammatory pathway, and environmental carcinogens, such as HCV and aflatoxin, to cause HCC development. HBV is DNA virus capable of integrating into the host DNA. This represents a crucial step in the pathogenesis of HBV-related HCC. Integration can occur at multiple or single sites and the inserted DNA is always alters[16]. Insertion of viral DNA leads to a series of genetic changes, including indels, repeats, and chromosomal translocations. Integration of HBV DNA may also alter transcriptional activity of important cellular genes[13]. Additionally, the HBx gene is often an HBV integrant that remains functionally active, inducing cellular changes[31]. HBx promotes cell cycle progression, and inhibits growth regulators and tumor suppressor genes, namely p53[17-20]. *TP53* and *RB1* have both been associated with poor prognosis in patients with HBV-associated HCC[16,21]. This phenomena have also translated to ethnic disparities as *TP53* and *RB1* are mutated at a much higher frequency in Asian Americans than in European Americans (TP53: 43% vs. 21%; RB1: 19% vs. 2%)[22]. HBx protein causes regional hypermethylation of DNA that causes silencing of tumor suppressor genes, or global hypermethylation that leads to substantial chromosomal instability[17,19]. Given the role of HBV and its interaction with other HCC features in HCC development, it is important to understand how collective exposure to HBV and other risk factors contributes to tumor heterogeneity and disparities in clinical outcomes.

1.1.2.2 HCV infection

While HBV predominates over HCV as major cause of HCC in low and medium human development index (HDI) regions, HCV is a major cause in high HDI countries such as the US[23]. More than 120 million people in the world are chronically infected with HCV[13]. Almost all cases of HCV occur in a cirrhotic or fibrotic liver. HCV is an RNA virus that typically leads to persistent infection, leaving individuals susceptible to hepatic fibrosis, cirrhosis, or HCC[13]. The mechanistic

underpinnings of HCV infection to HCC remains largely unknown. Increasing evidence suggests the complex interplay between inflammatory response, gene activation, and viral clearance creates a tumorigenic environment that promotes genomic instability, resulting in HCC[24]. Moreover, HBV/HCV coinfection leads to more severe liver disease and an increased risk of liver cancer[25]. As the incidence of HCV continues to rise in the US, and co-infection with HBV will inevitably increase, there is a greater need to identify genetic differences in patients with HBV, HCV, or HBV/HCV-associated HCC, and their interaction with other HCC features to improve outcomes.

1.1.2.3 Aflatoxin

One of the most common food-borne risk factors for HCC is aflatoxin, accounting for an estimated 5-28% of HCC cases[26]. Aflatoxins are naturally occurring toxins produced by *Aspergillus flavus* and *Aspergillus parasiticus* that are commonly found on staple foods, particularly as wheat, maize, ground nuts, and rice[27]. It is estimated that over 5 billion people globally have been exposed to these toxins[27]. Tropical and subtropical areas of the world have the highest rates of contamination, mainly sub-Saharan Africa, Eastern Asia, and parts of South America[28]. Exposure causes hepatotoxicity, teratogenicity, and immunotoxicity, leading to serious long-term health threats. Aflatoxin is metabolized in the liver by cytochrome p450 enzymes into an intermediate that is highly reactive and forms derivatives with other molecules, and can react with the p53[29]. As a result, aflatoxin binds to DNA generating an aflatoxin pro-mutagenic adduct which can then be converted to a stable adduct and lead to transversion mutations[29].

1.1.2.4 Metabolic syndrome

The incidence of obesity continues to climb at alarming rates, especially in the US. In most cases, obesity can be attributed to the metabolic syndrome (MetS)[30]. Metabolic syndrome is defined by having three or more of the following cardiometabolic risk factors: the presence of

obesity (body mass index (BMI) > 30 kg/m^2 or increased waist circumference), insulin resistance, dyslipidemia, and hypertension[31]. Paralleling this increase in obesity is the increase in incidence of HCC, and the growing association between the burgeoning rates of obesity and the striking increase in HCC tumors among the patients in the US[30]. It is estimated that 34% of the adult population (32.2% in males and 35.5% in females) in the US currently meet the diagnostic criteria for MetS[32].

Among patients with non-alcoholic fatty liver disease (NAFLD), MetS is strongly associated with features of NAFLD[33]. NAFLD is defined by presence of steatosis in >5% of hepatocytes, chronic alcohol abuse and viral hepatitis[34]. The risk of steatosis is greatly higher in individuals with MetS than individuals without non-MetS, and they're at risk of developing cirrhosis and HCC. NAFLD is responsible for between 5%-20% of HCC cases in the US[33]. Individuals with non-alcoholic steatohepatitis (NASH) and cirrhosis are at an increased risk of developing HCC, as high as 5.29 per 1,000 persons-years[35]. NASH is quickly becoming one of the most common risk factor for HCC, with 45% of cases developing in the setting of NASH[36]. While progress has been made to understand these risk factors, much remains to be learned about the mechanisms by which metabolic syndrome and associated features cause HCC. It is clear that NASH, NAFLD, and other alterations to normal liver biology and function play central roles in HCC development and progression. In this work, we will use pathway level analysis to examine liver biological function.

1.1.2.5 Alcohol overconsumption

Alcohol abuse is more common in the US and Western Europe, and is rising in Asia[7]. Over 15 million people in American abuse alcohol, increasing risk for the development of HCC. Chronic alcohol use (>80 g/day) for more than 10 years increases the risk of HCC five-fold[37]. Alcohol abuse accounts for 32-45% of HCC cases in the US[37]. Alcohol overconsumption significantly increases the risk of developing HCC in patients with HBV or HCV infection as

compared to patients with either viral infection who did not abuse alcohol[37,38]. Chronic alcohol abuse alters genomic integrity and liver structure, leading to cirrhosis, fibrosis, and eventually, development of HCC.

Collectively, hepatocellular carcinoma poses a significant health concern for Black people, Hispanics, and individuals > 65 years of age. HCC has a complex set of established risk factors, including lifestyle, occupational and environmental factors. Our work will focus on elucidating how these factors interact with other pathways to contribute to the initiation and progression of HCC.

1.1.3 Diagnosis, prevention, and treatment

HCC is most often diagnosed at advanced stages. The stage at diagnosis plays a crucial role in determining treatment management. Early diagnosis and efficacious treatment modalities remain a challenge. Diagnosis of HCC is typically based on non-invasive criteria, but there is increasing need for molecular markers for use in the clinic as some patients can be asymptomatic[4,39]. Aspirin and coffee have been implicated in preventative measures beyond HBV vaccination and anti-viral therapies for HBV and HCV[40]. The treatment management of HCCs have greatly improved over the last decade[4]. The prognosis is significantly better when detected early, where the primary curative option is hepatic resection or liver transplantation[41]. Despite alternatives, local ablation with radiofrequency is the mainstay of image-guided ablation for nonsurgical early-stage HCC[42,43]. Transarterial chemoembolization (TACE) is the most commonly used treatment and standard of care for intermediate-stage HCCs[41]. Advances in patient stratification have improved surgical outcomes and the 10-year survival rate post-transplantation[1,42]. However, for the considerable majority of HCC patients, their tumors are detected at an advanced stage where surgical cure is no longer viable[1]. Most patients, approximately 50-60%, will require systemic therapy, including immune-checkpoint inhibitors (ICIs), tyrosine kinase inhibitors (TKIs), and monoclonal antibodies. Currently, the most common

systemic therapy is sorafenib, a multikinase inhibitor against Raf, vascular endothelial growth factor receptor (VEGFR), and platelet-derived growth factor receptor (PDGFR) kinases[44]. The combination of atezolizumab (anti-PDL1) and bevacizumab (anti-VEGF) has more than doubled survival and improved patient-reported outcomes in advanced-stage HCCs[45]. Presently, sorafenib and lenvatinib remain the most effective single-agent drugs. Although targeted therapeutic agents, sorafenib and lenvatinib only modestly improve patient survival by 7-10 months[46-48]. Additionally, the single-agent therapies regorafenib[49], cabozantinib[50], and ramucirumab[51] have shown improved survival outcomes. While single-agent ICIs are proven to have remarkable clinical benefits in 15-20% of responders, identifying this group remains an issue as biomarkers have failed to detect groups[52,53]. Trials investigating combinatorial therapies, ICIs with TKIs or PD1/PDL1 with CTLA4 inhibitors, are in progress. These findings are projected to change HCC management at all stages. Given the success of moderate success of immunotherapies in HCC, further understanding the complexities of the HCC microenvironment has great potential to improve treatment management. To address this unmet need, this dissertation will further characterize signaling pathways and molecules implicated in HCC to identify factors influencing prognosis.

1.2 Mechanisms/pathophysiology

The pathophysiology of HCC is a complex stepwise process. The cooperation of several factors is responsible for the initiation of hepatocyte malignant transformation and HCC development. These elements include genetic predilection, joint interactions between viral and non-viral risk factors, the cellular and immune microenvironment, and the severity of chronic liver disease. The dysregulated microenvironment is vital to facilitating carcinogenesis, and it is well established that it participates in all stages of tumor progression. The cell of origin for HCC is controversial with much debate around the presence and role of stem cells in liver itself. The cell of origin could potentially be a hepatic stem cell, a transit amplifying cell or a mature

hepatocyte. Mature hepatocytes are hardy cells capable of retaining remarkable proliferative and regenerative capacity in response to liver injury. Multiple mouse studies substantiate that HCC originates in differentiated mature hepatocytes, while others purport that liver stem cells may be the source[54,55]. Paradoxically, a subset of hepatocellular carcinoma tumors display mixed HCC and CCA morphology, suggesting they seem to arise from mature hepatocytes and emphasizing the concept of transdifferentiation[54,56]. These observations underscore that phenotype and epigenetic landscape of tumors may not always directly reflect the cell of origin. We will study genomic alterations and dysfunction with meaningful prevalence and relevance to transdifferentiation of hepatocytes.

1.2.1 Hepatocellular genetics

Advancements in sequencing technologies have enabled the identification of key cancer genes in HCC. Mutations in *CTNNB1* (encodes the β-catenin protein), *AXIN1* or *APC* (inhibitors of WNT signaling) lead to the activation of WNT/β-catenin signaling pathway seen in 30-50% of HCC cases[12,57]. Other frequently altered mutations involved in cell cycle control are *TP53, RB1, CCNA2, CCNE1, PTEN, ARID1A, ARID2, RPS6KA3* or *NFE2L2*. Also, gene variations in molecules involved in epigenetic regulation, oxidative stress, and the AKT-mTOR and MAPK pathways have been linked to HCC progression[58]. Moreover, frequent focal chromosome amplifications in *CCND1*, *FGF19*, *VEGFA*, *MYC* or *MET* result in the activation of oncogenic pathways, including tyrosine kinases[59]. Although, HCC is marked by few recurrent mutations and many occur randomly, there are specific genes and pathways associated with HCC molecular class, outlined in **section 1.2.3**. Based on standards, only 20-25% of patients have at least one possible actionable target[12,39,60].

Another important facet is how risk factors play a role in HCC genomics. A prime example of this is that HBV infection potentiates the toxic effects of aflatoxin, specifically in patients with null polymorphism of *GSTT1*[61,62]. HBV-mediated mutagenesis is most frequently

mutated within the TERT promoter, resulting in overexpression of telomerase[20]. Aberrant activation of telomerase keeps cells from senescence and promotes cell transformation[63]. Other recurrent insertions associated with HBV participate in cell cycle control to activate oncogenes, such as *CCNA2* or *CCNE1*. Complex rearrangements and replicative stress are induced by these oncogenic alterations[64]. Collectively, these observations emphasize that viral infection activates specific oncogenes that function as early mediators of hepatocyte transformation. In contrast, HCV infection does not appear to promote direct oncogenic effects and the induction of mutations in HCV-related HCCs results from oxidative stress triggered by inflammatory response. Additionally, the severity of NASH and HCC incidence has been identified to be associated with polymorphisms in *PNPLA3*, *TM6SF2* and *HSD17B13*, particularly in the setting of chronic alcohol abuse[62,65].

1.2.2 HCC-related mutational signatures

While much of DNA sequencing has focused on recurrent driver mutations, a new approach called mutational signatures allows the analysis of all the genomic aberrations which reflect the underlying mutational processes and dysregulated events during carcinogenesis. Mutational signatures are a result of genetic and environmental stimuli that leave a distinct fingerprint of the acquired DNA damage. The ICGC/TCGA Pan Cancer Analysis of Whole Genomes (PCAWG) Network[66] have reported 49 mutational signatures using data from more than 23,000 cancer patients. Mutational signatures are pertinent to overall cancer biology, as they are detected in all cancer types, irrespective of the driver mutations, and appear to be associated with underlying etiologically relevant cellular processes and exposures[66-69]. Mutational signatures reflect genomic instability and DNA repair defects that play a role in molecular vulnerabilities, including chemotherapy response, and as such, mutational signatures can provide insights into mechanisms underlying disparities in outcomes. Throughout the development of cirrhosis and chronic liver disease, hepatocytes increasingly acquire numerous

genetic modifications and epigenetic changes. Several risk factors involved in inducing DNA mutations are connected with specific mutational signatures[12,66,68,70]. Exome sequencing of HCC identified single base substitution (SBS) signatures 4 and 29 with tobacco smoking and tobacco chewing, respectively [12,70]. Interestingly, the alcohol associated signature 16 exhibited transcription coupled damage and was responsible for majority of *CTNNB1* mutations which is linked with chronic alcohol consumption[12,70]. Signatures 22 and 24, notably in patients of Asian and African descent, were associated with aristocholic acid and aflatoxin B1, respectively[12,70-72].

It has been increasingly demonstrated that mutational burden and genomic instability of cancer are important for chemotherapeutic response[73]. Several mutational signatures are associated with distinct clinical outcomes and have emerged as potential biomarkers for novel targeted therapies[70,74-78]; many DNA repair pathways have been shown to represent molecular vulnerabilities that dictate chemotherapeutic response[79]. Some of these differences may be targetable. Previous studies have demonstrated that overexpression of DNA damage repair pathway confers resistance to radiotherapy and cisplatin[80,81], and for tumors with HRD and/or BRCA1/2 mutations, the use of PARP inhibitors which target compensatory DNA repair pathways, have been shown to be extremely efficacious in overcoming chemoresistance in both *in vitro* and *in vivo* HCC models[82-84]. The potential for these observations to be translated in the HCC preventative and management remains to be seen. Nonetheless, these signatures highlight the importance of normal liver function, which when dysregulated damages hepatocytes leading to carcinogenesis.

1.2.3 HCC molecular classes

Previous studies have established molecular and immune classes defined based on main molecular drivers and pathways or immune status that correlate with specific genomic features, histology, and clinical outcomes[11,39,58,85-90]. HCC can be characterized by two main molecular groups based on transcriptomics, proliferative and non-proliferative class[1,11,59,86,88-90].

The proliferation class, accounting for ~50%, is characterized by more clinically aggressive tumors with poor histological differentiation, high vascular invasion, and increased levels of α-fetoprotein (AFP)[91]. This group can be further separated into two subclasses: S1 or iCluster 3 and S2 or iCluster 1[11,88]. S1/iCluster 3, which represents 20% of HCC cases, are marked by Wnt–TGFβ activation, which drives an immune-exhausted phenotype[87]. Whereas S2/iCluster 1, representing 25-30% of HCC cases[59,88], are marked by progenitor-like characteristics with the expression of stem cell markers (*CK19*, *EPCAM*) and activated IGF2 and EPCAM signaling pathways[91]. There is an enrichment of HBV-related tumors present that activate classical cell cycle pathways, such as PI3K-AKT-mTOR, RAS–MAPK, MET and IGF pathways. Other key hallmarks of this class are high chromosomal instability, frequent *TP53* mutations, amplifications of *FGF19* or *CCND1* and global hypomethylation. Conversely, the non-proliferation class[11,59,88-90] is distinguished by less aggressive tumors with hepatocyte-like histological differentiation, decreased vascular invasion, and lower levels of AFP[91]. Tumors in this class are associated with NASH, NAFLD, and HCV infection. Unique subclasses have been characterized within this class: the WNT/β-catenin *CTNNB1* and the interferon subclass. The WNT/β-catenin CTNNB1 subclass presents frequent *CTNNB1* mutations, activation of the WNT/β-catenin signaling cascade, and extensive promoter hypermethylation (*CDKN2A, CDH1*), driving an immune-excluded phenotype with low immune infiltration[59,87,90]. Samples classified as immune-excluded are suggested to be mostly resistant to ICIs[92]. The interferon subclass shows a very activated IL6-JAK-STAT pathway with a more inflammatory tumorigenic microenvironment (high T cell infiltrate)[87]. This subclass also displays chromosomal stability with frequent TERT promoter mutations[86].

Most studies agree on a two-classification schema for DNA repair groups, characterized by low or high repair expression/activity[93-96]. These groups found high repair associated with poor prognosis, and further linked the classes with distinct immune profiles, poor differentiation,

elevated intratumor heterogeneity and mutation burden. One other study, using a manual clustering analysis, identified four groups with similar outcomes based on repair expression. However, the combined effect of DNA repair, liver regeneration, and other HCC features on prognosis remains unclear. Pathway level signatures represent a method to examine molecular interactions occurring in HCC and interrogate if these genes may predict outcomes. This will be crucial for understanding heterogeneity in HCC biology and clinical outcomes.

1.3 DNA repair mechanisms

DNA damage has long been established as a causal factor for cancer development. DNA repair pathways are activated to protect genetic stability and integrity when cells are exposed to DNA-damaging agents. Dysregulation of DNA repair pathways facilitates the initiation and progression of cancer through malignant transformation of cells. In addition, understanding DNA repair dysfunction has the potential to contribute valuable information for HCC clinical decision making. Considering the lack of diagnostic and predictive biomarkers in HCC, the development of molecular markers will fill a substantial need in the clinical setting.

1.3.1 DNA repair pathways and their role in cancer

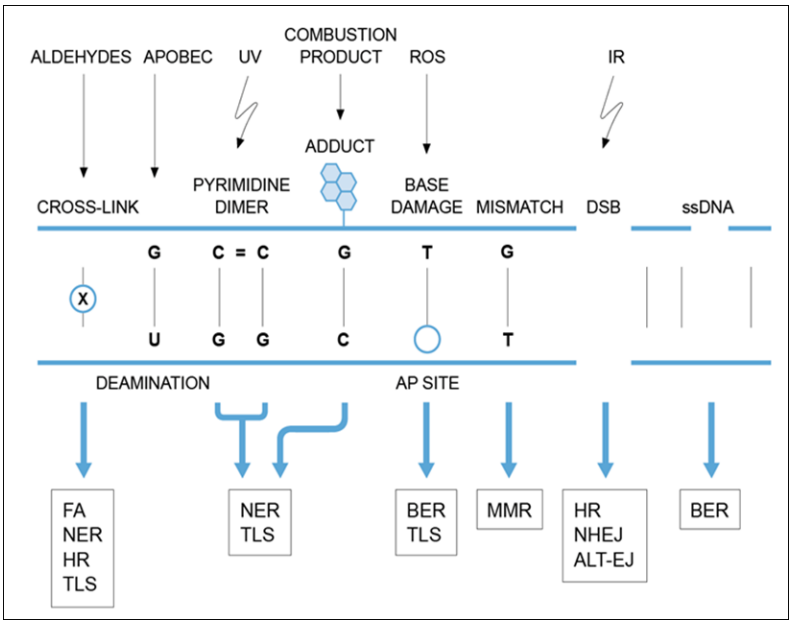
DNA repair systems are classified into the following major pathways: 1) direct reversal, which mainly repairs the lesion induced by alkylating agents, 2) nucleotide excision repair (NER), correcting bulky, helix-distorting DNA lesions, 3) base excision repair (BER), aiming at DNA breaks (SSBs) and non-bulky impaired DNA bases, 4) recombinational repair, which is further divided into homologous recombination (HR) and non-homologous end joining (NHEJ), primarily functioning at DNA double strand breaks, 5) mismatch repair (MMR), involved in replication errors, 6) alternative nonhomologous end joining (alt-NHEJ), involved in repair of DSBs, 7) translesion synthesis (TLS), which is more likely to be a DNA damage tolerance mechanism[97,98] (**Figure 1.1.1**).

Numerous studies have found that certain cancers are associated with defects or mutations in DNA repair pathways[66,75,98]. For example, two important HR-related genes, *BRCA1* and *BRCA2,* when mutated in the germline confer genetic predisposition for breast, ovarian, and pancreatic cancer[99]. Association between homologous recombination deficiency (HRD) and mutational burden have been previously demonstrated in breast[100], lung[101], and other cancers[102]. Likewise, *RAD51* is the central recombinase of the HR pathway, and therefore plays a role in maintaining genomic stability[103]. RAD51 is recruited in response to DNA double strand breaks or in the instance of replication fork stalling before or after DNA repair. DNA double-strand break generation and recognition results in nuclear accumulation of RAD51, which can be detected at the foci[103]. RAD51 is responsible for stabilizing single-strand DNA by filament formation, searches for homologous DNA regions and mediates strand exchange[104]. Surprisingly, *RAD51* mutations are rarely mutated which are associated with cancer predisposition[103]. Rather, overexpression frequently occurs in tumors, and is associated with worse prognosis[103,105]. Upregulation of RAD51 is associated with enhanced DNA repair and increased chemotherapeutic resistance[106]. Conversely, downregulation of *RAD51* was noted in multiple tumor types due to hypoxia, implying that that the hypoxic tumor microenvironment suppresses the HR pathway to produce genetic instability[107,108]. Despite this paradox, it is clear *RAD51* and other HR-associated genes have implications in carcinogenesis, cancer progression and anticancer drug resistance. Therefore, DNA repair dysfunction can contribute to tumor development by promoting genomic instability and aberrations. Assessment of DNA damage and repair dysregulation may pose another method of stratifying patients based on DNA repair status for improved treatment management and selection.

Figure 1.1.1 DNA damage and repair pathways.

1.3.2 DNA repair pathways and cancer therapeutic resistance

The leading cause of treatment failure is resistance to therapy. The primary anti-cancer therapies induce cell death by directly or indirectly causing DNA damage. DNA damage response dysfunction may contribute to hypersensitivity or resistance of cancers to genotoxic agents, and targeting the DNA repair pathway could increase tumor sensitivity to cancer therapeutics. For instance, it has been reported that the treatment of HCC is difficult as a result of properties in their DNA damage response (DDR)[109]. One of the most widely employed DNA-inhibitor drugs is Cisplatin. Cisplatin-resistant tumor cells exhibited greater levels of DNA damage repair related genes and DNA repair capacity, and inhibition of NER pathway substantially enhanced the sensitivity of cells to cisplatin[110,111]. In triple negative breast cancer (TNBC) patients treated with breast-conserving surgery and radiotherapy, low expression of



TP53BP1, a key protein in NHEJ, associated with higher local recurrence, suggesting that TP53BP1 may be a predictor of radio-resistance[112].

A number of DDR inhibitors have entered the market or are undergoing development. In this realm, poly(adenosine 5′-diphosphate) ribose polymerase (PARP) inhibitors are a group of targeted cancer drugs that work by blocking enzymes involved in DNA repair mechanisms that cancer cells relay on to multiply. PARP inhibitors, such as olaparib, niraparib, and rucaparib, have been widely used in patients with BRCA1/2 germline mutations or HRD[113,114]. The OlympiAD study showed olaparib produced better progression-free survival in patients with germline BRCA1/2 mutations, compared to chemotherapy[115]. However, this improvement is hindered by eventual resistance in the majority of patients leading to treatment failure[114]. Additionally, upregulation of oncogenic pathways such as WNT/β-catenin pathway or DDR related proteins may also confer sensitivity to PARP inhibitors, supporting evidence for combinatorial approaches with PARP inhibitors[97,116,117]. Combinatorial approaches are not without drawbacks as they may increase the risk of mutagenic lesions in surviving cells, leading to the development of secondary tumors. Also, the role of DDR in immunotherapy in garnered much attention. Researchers have shown that DNA repair deficiencies associated with immune checkpoint blockade (ICB) response. For example, MMR has been described as a biomarker of response to immune checkpoint inhibitors (ICIs)[118].

Given the central role of genomic instability in cancer, targeting DNA repair pathways poses a potential treatment approach for HCC tumors. Gaining a better understanding of the biological underpinnings and regulatory mechanisms of DNA repair pathways could facilitate development of modalities for enhancing anticancer effect of DNA-damage based therapies. Furthermore, the identification of DNA repair classes may aid in patient stratification and treatment management, especially in patients who are not eligible for surgical intervention.

1.4 Summary

Hepatocellular carcinoma is very heterogenous disease that can be classified into two major molecular classes, which have differing histology, molecular features, treatment, and

prognosis. These differences are likely a result of multiple pathway interactions but much remains unclear and efforts to expand beyond the two-classification schema is made difficult the normal liver regenerative environment. The proposed work will focus on characterizing and further subdividing HCC to gain insight into heterogeneity. Our work aims to elucidate factors interacting with molecular and cellular pathways to contribute to the initiation and progression of HCC. The complex set of established risk factors play a role in genomic integrity and liver function[58]. Specifically, DNA repair pathways have been linked to producing variability in clinical outcomes[89,119-121]. DNA repair damage and repair elicits specific mutational patterns, called mutational signatures. Mutational signatures represent a method to detect repair deficiencies associated with genomic instability and risk factors, making them ideal candidates for biomarkers of response. The potential of matching mutational signatures with DNA repair signatures could help identify class-specific markers and therapies. This is particularly important given diagnosis of HCC typically occurs at advanced stages, making treatment options limited. This project will allow better identification of intermediate and advanced tumors suitable for targeted therapies.

The objective of this proposed work is to characterize HCC, integrating across multiple platforms to evaluate how molecular and cellular pathways vary and interact to impact patient survival. In chapter 2, we characterize HCC subclasses utilizing an integrated TCGA HCC and CCA dataset to gain insights into biology. Chapter 3 extends this work to elucidate RNA-based DNA repair signatures. We evaluate mitotic and regenerative signatures, and clinicopathologic variables as mediators of HCC repair response and resistance. Our molecular classes and repair signatures have implications for targeted therapy, ICIs, and combination therapies. Advancements have been made in HCC treatment management, but identifying patients responsive to DNA repair targeted therapies and immunotherapies remains an issue due to lack

of biomarkers. The prognostic utility of single agent ICIs in HCC has only shown moderate improvements, as compared to the standard of care for HCC sorafenib.

Through investigation of repair groups as mediators of HCC resistance, it is possible that we may identify mechanisms by which intermediate to advanced HCCs may benefit from a targeted, immunotherapy, or combinatorial drugs. This information will address the need for improved biomarkers options for poor prognosis groups in HCC. Further understanding the genomic underpinnings of HCC and their interactions with liver signaling pathways may help define features stratifying prognostic groups and facilitate the discovery of targeted therapies to improve outcomes.

CHAPTER 2: GENOMIC CHARACTERIZATION OF RARE MOLECULAR SUBCLASSES OF HEPATOCELLULAR CARCINOMA[1]

2.1 Overview

Background: Primary liver cancer, consisting of both cholangiocarcinoma (CCA) and hepatocellular carcinoma (HCC), is the second leading cause of cancer deaths worldwide.

Methods: Our goal is to genomically characterize rare HCC subclasses to provide insight into disease biology. Leveraging The Cancer Genome Atlas (TCGA) to perform a combined analysis of CCA (n=36) and HCC (n=374), we integrated multiple genomic platforms, to assess transcriptional profiles, mutational signatures, and copy number patterns to uncover underlying etiology and linage specific patterns.

Results: We identified two molecular classes distinct from prototypical HCC tumors. The first, CCA-Like, although histologically indistinguishable from HCC, had enrichment of CCA mutations (*IDH1*, *BAP1*), mutational signatures, and transcriptional patterns (*SOX9*, *KRT19)*. CCA-Like, however, retained a copy number landscape similar to HCC, suggesting a hepatocellular linage. The second, Blast-Like, is enriched in *TP53* mutations, HBV infection, exposure related mutational signatures and transcriptionally similar to hepatoblasts. Although these subclasses are molecularly distinct, they both have a worse progression-free survival compared to classical HCC tumors, yet are clinically treated the same.

Conclusions: The identification of and characterization of CCA-Like and Blast-Like subclasses advance our knowledge of HCC as well as represents an urgent need for the identification of class specific biomarkers and targeted therapy.

[1]A version of this work has been previously published in Nature Communications Biology. The original citation is as follows: Damrauer, J.S., Smith, M.A., Walter, V. *et al.* Genomic characterization of rare molecular subclasses of hepatocellular carcinoma. *Commun Biol* **4,** 1150 (2021).

2.2 Introduction

Primary liver cancer is the 2nd and 6th leading cause of cancer death worldwide for males and females, respectively[122]. Within the United States, primary liver cancer rank as the 5th (males) and 7th (females) most deadly cancer[123]. Primary liver cancer includes both hepatocellular carcinoma (HCC) and cholangiocarcinoma (CCA), and although they are anatomically co-localized, they have different etiologic and genomic features[124,125]. HCC is thought to be derived from hepatocytes and accounts for 90% of all primary liver cancer. It has well characterized risk factors including chronic hepatitis B/C (HBV and HCV) infection, alcohol abuse, diabetes, and aflatoxin exposure[54]. CCA is the second most common primary liver cancer and stems from biliary cells. CCA risk factors include: primary sclerosing cholangitis, hepatobiliary flukes and biliary tract cysts[126]. Recent publications by The Cancer Genome Atlas (TCGA) identified *IDH1* and *IDH2* mutations, a common feature of CCA, in a subset of HCC samples[127,128]. These *IDH1/2* mutant tumors showed similar gene expression patterns observed in CCA based on ~2,000 genes as well as displaying similar methylation patterns as other *IDH* mutant tumors. This suggests that HCC tumors may be sub-classified based on their relatedness to CCA. Due to the dearth of targeted treatment options for HCC, further subdividing and characterizing HCC, particularly additional characterization of subset set similar to CCA, may aid in the understanding of the disease and, in the future, lead to the identification of new therapeutic targets.

Previous attempts to classify HCC tumors have identified patient populations with gene expression, mutational or survival differences, although few groups have done so with a combined HCC and CCA harmonized dataset; however, these prior studies were with a limited number of data types, small cohorts or within a singular ancestral or etiologic group[88-90,127,129,130]. Our work expands on the previous studies by using a large, harmonized cohort (CCA and HCC), with integrated multi-omic data of samples not restricted to any singular etiology.

Through a multi-omic approach utilizing a fully integrated CCA and HCC dataset, we define three distinct subpopulations of hepatocellular carcinoma tumors, CCA-Like, Blast-Like and HCC. We integrated these subpopulations with external datasets anchoring our data to lineage and cell type specific cells and indicating a derivation from a hepatocyte lineage.

2.3 Methods

2.3.1 Tumor Classification

Upper quartile normalized RSEM gene expression data for TCGA was downloaded from the GDC legacy archive (https://portal.gdc.cancer.gov /legacy-archive/). Cholangiocarcinoma (CCA dataset, n=36) and Hepatocellular carcinoma (HCC dataset, n=374) samples from TCGA were merged, log2 transformed, and filtered for highly expressed and variably expressed genes (n=4035). The data was median centered across genes. The Spearman correlation to the median gene expression of all CCA samples was calculated to determine the per sample similarity for all CCA and HCC samples. Samples within (-) 1 standard deviation from the mean correlation of all CCA samples were classified as CCA-Like. To determine similarity to hepatoblast cells, we used single cell RNAseq data of hepatoblast differentiation in mice (GSE90047)[131]. HCC samples were correlated to variably expressed genes from hepatoblasts (E10.5) and differentiated hepatocytes (17.5, DLK^+, $EPCAM^-$) and cholangiocytes (17.5, DLK^-, $EPCAM^+$). HCC samples with a correlation to hepatoblasts in the upper tertile of all samples and not otherwise classified as CCA-Like were classified as Blast-Like. The resulting classification yielded, CCA (n=36), CCA-Like (n=33), Blast-Like (n=66), HCC (n=275).

2.3.2 mRNA Analysis

All external datasets were log2 transformed and median centered across genes. The Spearman correlation value was calculated between the CCA/HCC samples and the median

expression of the comparison classes (CCA, HCC, Normal bile duct (NBD), Liver, Cholangiocytes, Hepatocytes and Hepatoblasts). To compare the CCA/HCC dataset to microdissected normal bile duct and normal liver, the TCGA dataset was merged with GSE26566 by adjusting TCGA data to the median expression of GSE26566 cholangiocarcinoma samples. The Spearman correlation was calculated as described above comparing the CCA and HCC cohort to microdissected normal bile duct, normal liver, and cholangiocarcinoma. To generate a differentiation score, the per sample correlation to Normal Liver was subtracted from that sample's correlation to NBD. This was also done for correlations to hepatocytes and cholangiocytes. Hepatitis B virus was detected in CCA and HCC tumors RNAseq data via VirDetect[132]. To visualize gene expression patterns across known markers of cholangiocytes and hepatocytes, CCA/HCC samples were hierarchically clustered (Cluster3.0), using expression markers from Hu et. al.[133]. Gene Set Enrichment (GSEA) was performed one vs rest comparisons across classes for all HCC tumors. Significance was determined using a nominal p-value <0.05 and FDR <0.25.

Hoshida and Woo subtypes were derived by extracting the respective gene signatures and performing ConsensusClusterPlus[134] to determine the expression groups. Hierarchical clustering was performed using the ES1 signature from Ben-Porath et. al., the samples within the increased gene expression cluster were selected as ES1 enriched[135].

Variant calling on RNA-Seq data was performed by aligning RNASeq reads with STAR[136] in two pass mode with unmapped reads assigned to the mate's position when possible. Parameters outFilterScoreMinOverLread and outFilterScoreMinOverLread were set to 0.45. Reads were realigned using ABRA2[137]. Reads were sorted and duplicates were marked using biobambam[138] both before and after running ABRA2. Indels were called using Cadabra. All analysis was performed in R (Version 3.5.2) unless otherwise noted.

2.3.3 miRNA Analysis

RSEM data was downloaded from the GDC legacy archive (https://portal.gdc.cancer.gov/legacy-archive/search/f) and log2 transformed. To determine significantly differentially expressed miRNAs, t-tests were performed on a per gene basis, CCA-Like vs. Blast-Like/HCC and CCA-Like vs. CCA. Benjamini-Hochberg adjusted p-values were calculated to account for multiple comparisons.

2.3.4 Genomic Features

Copy number and mutation data was downloaded from FireBrowse (http://gdac.broadinstitute.org). Lollipop plots were generated through cBioPortal.org[139,140]. Additional *BAP1* alterations were determined using the de novo aligner ABRA2[137]. Chi-square and Fisher's exact test were performed when appropriate in a pairwise manner. For mRNA expression, a two-sample t-test was performed to determine significance of expression with CCA/HCC classification classes. The biomaRt R package[141] was used to identify genomic positions for genes. Custom R scripts based on functions in the MVisAGe R package[142] were used to plot mean gene-level DNA copy number values in each expression subtype as well as differences of mean gene-level DNA copy number values between pairs of gene expression subtypes.

To identify disease class specific copy number alteration (CNA) we used SWITCHplus (https://genome.unc.edu/SWITCHplus/)[143]. Fisher's exact test was performed between paired comparison classes to identify class specific CNAs. Segment's significance was assigned using Benjamini-Hochberg adjusted p-values <0.05.

2.3.5 Characterizing Mutational Signatures in HCC and CHOL

The R package SomaticSignatures was used to identify mutational signatures in 396 TCGA HCC and CHOL whole exome sequencing samples[144]. Motif contributions across the samples were aggregated by RNA class. The R package barplot3d[145] was used to generate 3D barplots displaying the frequency of the 96 different combinations of somatic mutations and trinucleotide contexts seen in this cohort. Using COSMIC mutational signatures version 3[66], we performed cosine similarity (*CS*) between our six signatures and the 49 Single Base Substitution (SBS) signatures to further characterize the mechanisms underlying our signatures. Liver sample motifs were correlated with our signatures to determine similarities and whether signatures are subtype specific, as we previously saw with the correlation to COSMIC mutational signatures v3.

2.3.6 Class Prediction, Survival Analysis, And Clinical Variables

Using the CCA/HCC as the training dataset, a gene classifier (n=150) was generated using ClaNC[146]. Predictions were made on GSE14520 using the correlation to the training set centroids, and Kaplan-Meier curves were generated using the survminer package in R[147]. Univariate Cox proportional hazards models were used to determine the significance of the tumor classes, stage and grade. Variables significant in the univariate analysis were then incorporated into a multivariate model. Clinical data for TCGA was obtained from Liu et. al.[148]. Clinical data that was included in the analysis had >94% of data present across the cohort (race, gender, age, stage, grade, and survival). As part of TCGA, diagnostic and frozen slides were reviewed by panel of pathologists with expertise in hepatobiliary cancers. The panels consisted of 6 pathologists for HCC[127] and 5 pathologists for CCA[128].

2.3.7 Statistics and Reproducibility

Categorical variables were compared using Fisher's exact or chi-square test. Continuous variable comparisons were made using t-test or ANOVA as indicated. Correlations were performed using Pearson or Spearman correlation as indicated. Multiple comparison correction was performed using Bonferroni correction. Survival analyses were performed using Kaplan-Meier with log-rank tests. Statistical analyses were performed using R unless otherwise noted.

2.3.8 Data and Code Availability

TCGA samples were collected through an IRB-approved protocol or through a TCGA-specific IRB waiver. Informed consent and IRB approval was obtained for all other data by the authors of the publications in which the original datasets were published. TCGA data is available through the GDC data portal, https://portal.gdc.cancer.gov/. Expression data is available by download from https://www.ncbi.nlm.nih.gov/geo/ accession number: GSE14520 (Roessler et. al.[149]), GSE90047 (Yang et. al.[131]), GSE26566 (Andersen et. al.[150]). All underlying data for figures available here: https://doi.org/10.6084/m9.figshare.15180810.v1. All code is available upon request.

2.4 Results

2.4.1 A class of hepatocellular carcinoma tumors show cholangiocarcinoma gene expression patterns

To determine the similarity between TCGA hepatocellular carcinoma samples (HCC) and TCGA cholangiocarcinoma samples (CCA), we calculated each sample's correlation to a defined CCA centroid. Thirty-three HCC samples were highly correlated to the CCA centroid (CCA-Like > CCA mean - 1 S.D, mean=0.75, S.D. = 0.09) and were classified as

cholangiocarcinoma-Like (CCA-Like) (**Figure 2.7.1a**). The rest of the HCC samples had lower correlations to CCA, in a similar range as the tumor adjacent normal tissues.

As TCGA data comes from bulk specimens, we compared the CCA and HCC cohort to microdissected normal bile duct, normal liver, and cholangiocarcinoma from Andersen et. al. (**Figure 2.7.1b**). The CCA-Like tumors had a higher correlation to normal bile duct than normal liver and in similar range as the CCA samples from TCGA and Andersen cohorts. Whereas HCC samples more closely resembled normal liver, though with a larger range of correlation.

We further anchored the TCGA data to single cell RNA sequencing data derived from fetal murine livers (embryonic day E10.5-17.5), representing hepatoblasts (E10.5), cholangiocytes (E17.5) and hepatocytes (E17.5) (**Figure 2.7.1c**). The CCA and CCA-Like samples were correlated to both cholangiocytes and hepatoblasts. While most HCC samples were only correlated to hepatocytes, we also identified a class of HCC samples with high correlation to the hepatoblast cells that had not been previously classified as CCA-Like, these sample were classified as Blast-Like (n=66).

2.4.2 CCA-Like and CCA share genomic alterations

To further dissect the molecular and clinical characteristics of these tumors, we assessed a series of clinical variables (**Supplemental Table 2.8.1**) as well as mutation, copy number and gene expression markers of classical alterations in both the CCA and HCC tumors (**Supplemental Table 2.8.2**) (**Figure 2.7.2a**) [127,128]. As previously described, a subset of TCGA HCC samples had canonical *IDH1/2* mutations (p.R132C/p.R172), a known hallmark of CCA [127]. Interestingly, those mutations were almost exclusively found in the CCA-Like class, except for one HCC tumor with a DNA and RNA variant allele frequency of <0.1% and <0.001% respectively (**Supplemental Figure 2.8.1a**). Additionally, *IDH1* gene expression was significantly reduced in the CCA-Like tumors compared to the Blast-Like and HCC tumors

(**Supplemental Table 2.8.2**). *BAP1* has previously be shown to be frequently altered across both CCA and HCC[127,128]; however, it is almost universally altered in CCA. In the CCA-Like, *BAP1* mutation rate was similar to CCA and copy number levels were lost significantly more than in HCC, yet not to the same degree as CCA (**Supplemental Table 2.8.2**). As compared to HCC, CCA-Like had a decreased mRNA and protein expression of *BAP1* (**Supplemental Table 2.8.2**).

We assessed whether these shared features are due to the CCA-Like class representing a mixed hepatocholangiocarcinoma phenotype. TCGA's pathology re-review identified only 7 cases of hepatocholangiocarcinoma in the TCGA HCC cohort, five of which were in the CCA-Like class, representing 15% of CCA-Like samples (**Supplemental Table 2.8.1**) the remain CCA-Like samples were unambiguously classified as HCC (**Supplemental Figure 2.8.1b-e**). A recent study reported that approximately 8% of mixed tumors have *ARID1A* mutations[151]; however, we did not observe any *ARID1A* mutations in the CCA-Like group, while the other classes had mutation frequencies between 8-17%.

CCA-Like was almost devoid of the prototypical hepatocellular carcinoma mutations, *CTNNB1* and *TP53* when compared to Blast-Like, and this group exhibited significantly higher mRNA expression of *p53* when compared to HCC. Interestingly, the Blast-Like class had a significantly higher rate of *TP53* mutation (**Figure 2.7.2a**), specifically truncating mutations and R249S mutation (**Supplemental Figure 2.8.2**). As HBV is a risk factor for HCC, we wanted to identify tumors with concurrent HBV infections. Unfortunately, not all samples had the corresponding clinical annotation; therefore, we identified tumors that contained RNAseq reads corresponding to the HBV genome. Blast-Like tumors had increased rates of HBV infection (**Supplemental Figure 2.8.3a**) as well as disproportionately high number of patients with Asian ancestry (**Supplemental Table 2.8.1**) (**Figure 2.7.2a**). Regardless of tumor class, samples

from Asian individuals, had a significantly higher number of HBV mRNA reads (t-test p-value = 1.8e-35) compared the CCA-Like/HCC samples.

2.4.3 CCA-Like tumors have shared gene-expression features of CCA tumors.

The CCA/HCC cohort was clustered using a set of genes associated with either hepatocytes, biliary/progenitor cells or markers of the cell cycle identified from organoid studies in Hu et. al. (**Supplemental Figure 2.8.4a**) [133]. HCC tumors clustered alongside the tumor adjacent normal samples and had increased expression for hepatocyte markers such as *ALB* and *HNF4A*. CCA-Like tumors co-clustered with the CCA tumors and similarly had higher expression of the cell cycle and biliary markers but lower expression of hepatocyte markers. CCA had significantly reduced expression of the hepatocyte marker *HNF4A* (p=1.7e-10) (**Figure 2.7.2b**) and ALB (p=1.0e-10) (**Supplemental Figure 2.8.4b**) and increased expression of cholangiocyte marker *SOX9* (p=3.6e-21) as compared to HCC tumors (**Figure 2.7.2c**). While the CCA-Like were very similar to the CCA tumors, the CCA-Like cells demonstrated higher gene expression of hepatoblast marker *AFP* (p=8.6e-9) (**Figure 2.7.2d**) [152]. The Blast-Like tumors had increased expression in cell cycle markers, while displaying intermediate expression of both biliary and hepatocyte markers, with the exception of *AFP*, which was significantly higher in the Blast-Like tumors compared to all other classes (p<0.001 for all pairwise comparisons (**Figure 2.7.2b-d**, **Supplemental Figure 2.8.4a**) (**Supplemental Table 2**). Additionally, the Blast-Like tumors also had increased expression based stemness index, mRNAsi, compared to the other subtypes (Supplemental Figure 4c)[153].

We evaluated immune cell patterns across the classes. We visualized the Bindea gene signatures[154] representing 24 immune cell types and found that samples with high expression of any immune signature were generally high for all immune signatures (**Figure 2.7.2e**). Samples were grouped by overall median immune gene signature expression and we found the high

immune group was associated with a lower tumor purity and an increased DNA methylation leukocyte faction score (**Figure 2.7.2e**). CAA and CCA-Like classes were enriched with immune signatures and grouped into the immune high set compared to HCC class (**Figure 2.7.2f**).

2.4.4 Shared mutational motifs and mutational signatures between CCA and CCA-Like

Reproducible patterns of single nucleotide variants (SNVs), termed 'mutational signatures', give a snapshot of the mutational pressure cells have undergone, many of which associate with known mutagens[66,67]. We examined our classes to determine if mutational patterns differ by class, particularly in the context of liver cancer which is associated with exposures with known mutational signatures (e.g., aristolochic acid, aflatoxin, and tobacco). CCA-like tumors shared a similar enrichment of C>T/G>T mutations with the CCA tumor class (**Figure 2.7.3a-b**), while Blast-Like and HCC class tumors shared similar mutational patterns, with decreased C>T/G>T frequency and increased A>T/T>A frequency (**Figure 2.7.3c-d**). All 96 mutation contexts were hierarchically clustered to visualize the relationships among classes, and was consistent with our transcriptional findings that CCA-like is more related to CCA tumors (**Supplemental Figure 2.8.5a**). While the top three motifs (all nC>Tn) are shared across the 4 classes, they are most abundant in CCA and CCA-Like. Globally, CCA and CCA-Like also have more diversity overall of mutation motifs as compared to Blast-Like and HCC (**Supplemental Figure 2.8.5a**). The per sample motif patterns were then compared against a list of previously discovered single base substitution (SBS) signatures from the COSMICv3 database[66]. The median cosine similarity (CS) for each signature within subclass was calculated and hierarchical clustering was performed and visualized alongside the per sample values. As with the motif level comparisons, CCA and CCA-Like were the most similar with Blast-Like and HCC sharing common feature sets (**Supplemental Figure 2.8.5b**). We wanted to identify *de novo* mutational patterns. Six mutational signatures (S1-S6) were identified, with S1, S2, and S4 each being primarily driven by a single motif, T>A-CTG, T>C-ATA C>A-GCC, respectively (Supplemental

Figure 5c). Signatures S3 and S5 both were driven by the presence of C>T: GCG, CCG, ACG motifs, with S5 having a low-level increase in broader range of additional C>T motifs (Supplemental Figure 5c). The S6 signature, which lacked the presence of any one given motif at a high frequency, had some shared motif patterns with S2 and S5. We quantified the median contribution of each signature for each tumor class and found statistically significant differences between the tumor classes (**Figure 2.7.3e, Supplemental Table 2.8.4**). Signatures S3 and S5 composed the majority of contribution to CCA-Like, a shared feature with CCA, along with having decreased contributions of both S1, S2 and S4 (**Figure 2.7.3e**). There was no significant difference between CCA and CCA-like for signature S5; however, both classes were significantly enriched compared to Blast-Like (CCA p=2e-11, CCA-Like p=7e-5) and HCC (CCA p=2e-15, CCA-Like p=2e-6). CCA-Like had an increased prevalence of S6, as compared to CCA, a feature shared with the Blast-Like and HCC. Blast-Like and HCC displayed remarkable similarity to each other, with the exception of S4 (p=0.009), which was defining feature of Blast-Like tumors.

To identify possible etiologies of these signatures, we correlated the motif signatures to the COSMICv3 database (**Figure 2.7.3f**). Signatures S1 and S6 were highly correlated to mutational patterns of chemical or environmental exposures: chemotherapy treatment (S1, SBS25 Cosine Similarity (CS)= *CS*=0.72), aristolochic acid (S1, SBS22 *CS*=0.96), tobacco (S6, SBS29 *CS*=0.83) and aflatoxin (S6, SBS24 *CS*=0.74). Defects in mismatch repair patterns defined signature S2 (SBS6 *CS*=0.81, SBS15 *CS*=0.79) except for signature SBS26, which defined signature S3 (*CS*=0.73). Base excision repair (BER) defects dominate signature S5, specifically signatures derived from tumors with *NTHL1* (SBS30, *CS*=0.71) and *ERCC2* (SBS5, *CS*=0.79) mutations; however, these mutations were not observed within our CCA or CCA-like samples where this signature was enriched. Interestingly, signature S4 defined by SBS16 and enriched in Blast-like currently has an unknown etiology.

Signatures S1 and S6, which contribute most to the liver specific subtypes (Blast-Like and HCC, and to a lesser extent CCA-Like), have high correlations to known liver carcinogens, aristolochic acid and tobacco/aflatoxin, respectively. *TP53* mutations, specifically the R249S mutation, has been previously linked to aflatoxin exposure. We observed that tumors with the *TP53* R249S mutation had a significantly higher cosine similarity to the aflatoxin signature (SBS24) than either tumors with alternative *TP53* mutations (p=0.001) or tumors WT for *TP53* (p=3e-4) (**Supplemental Figure 2.8.5d**), similar to what had been observed in TCGA[127]. The Blast-Like class was dominated by signature S4, which although highly correlated to SBS16 (*CS*=0.74), it currently has no known etiology. Conversely, S3 and S5, are shared across CCA and CCA-Like classes, which lack exposure related correlations, but are enriched for mismatch repair signatures. These results suggest that the mutational pressures in the CCA-like are more similar to CCA and potentially highlighting different selective pressures than the more predominant exposure-based pressures observed in HCC and Blast-Like.

2.4.5 Transdifferentiation pathways are upregulated in CCA-Like tumors

We explored the genomic/transcriptomic features driving each tumor class to determine if there are signaling pathways that are shared across classes. We performed gene set enrichment analysis (GSEA) using the hallmark gene signature list to identify differential pathway signatures between CCA-Like and the other HCC tumor classes (Blast-Like and HCC) (**Figure 2.7.4a**) [155,156]. Of the seven significant gene sets (nominal p-value <0.05 and FDR<0.25), three pathways, TGFβ, NOTCH and WNT have previously been implicated as drivers of Epithelial-Mesenchymal Transition (EMT) and transdifferentiation in the liver (**Supplemental Figure 2.8.6a-c**) [157-163]. TGFβ, NOTCH and WNT expression signatures were all significantly elevated within the CCA-Like tumors compared to the HCC classes (CCA-like vs HCC p<0.001) (**Figure 2.7.4b-c, Supplemental Figure 2.8.6d**). For these three signatures, the CCA-Like tumors module scores were not significantly different to those observed in CCA albeit

slightly reduced (TGFβ p=0.015, NOTCH p=0.70, WNT p=0.62). We further compared the CCA-Like tumors directly to CCA samples. Four of the top five significantly enriched pathways in CCA-Like tumors related to liver biology (**Figure 2.7.4d**). We plotted the median expression of 10 Cytochrome P450 (CYP) genes that are abundant in the liver as a surrogate for liver specific gene expression and found that CCA-Like tumors had significantly higher levels of expression of CYP genes compared to CCA tumors (t-test, p=3.23e-5). CCA-Like tumors expressed CYP at comparable levels to the Blast-Like tumors which was still lower than what was observed in HCC or adjacent normal liver tissue (**Figure 2.7.4e**).

Pathway level analysis identified multiple pathways that are involved in both transdifferentiation as well as EMT. miRNA expression has been shown to drive both processes. We performed a differential miRNA expression analysis and identified four mir-200 family members highly enriched in both CCA and CCA-Like tumors (**Figure 2.7.4f-g**) compared to Blast-Like and HCC. Blast-Like and HCC samples had increased expression of miR-122, a liver specific miRNA[164]. Interestingly, when CCA-Like and CCA are directly compared, miR-122 was the most enriched miRNA in the CCA-like, while the miR-200 family members were significantly enriched in the CCA (fold change (FC) = 3.7, FDR <0.001). Expression of one representative miR-220 family member, miR-200b-3p, was highest in CCA, followed by CCA-like and lowest in HCC (**Figure 2.7.4h**).

2.4.6 The CCA-Like copy number landscape resembles that of HCC

Although the CCA-Like class tumors bear a striking transcriptional resemblance to the CCA class, key markers (elevated *AFP/ALB* and miR-122) indicate that the CCA-like tumors still have features shared with HCC and suggests the precursor cell likely arises from hepatocytes rather than a cholangiocyte/hepatocyte progenitor cell. Because copy number alterations are

often considered early events in transformation[165], we compared the copy number landscapes of CCA, CCA-Like, Blast-Like and HCC to infer the shared cellular origin of the classes.

Overall, CCA-Like, Blast-Like and HCC displayed a more similar copy number landscape to each other than to CCA (**Figure 2.7.5a**). Using SwitchDNA, we performed a pairwise comparison of segments comparing CCA-Like to CCA and HCC, and Blast-Like to CCA and HCC. CCA-Like had significantly fewer copy number differences with HCC (n=360) than CCA (n=1024) ($p<0.0001$) (**Supplemental Figure 2.8.7a**). There were 272 segments that were significant in CCA-Like in both comparisons to CCA and HCC, all of which (except one segment) were between 3p24.3-12.3. Located within this region is *BAP1,* which is almost universally lost in CCA. *BAP1* was lost to a lesser extent in CCA-Like (45% of samples) than CCA (80% of samples); however, at a significantly greater frequency than HCC (12%) ($p<0.05$) (**Supplemental Table 2.8.2**, **Figure 2.7.5b**). Conversely, *FOXC1* (6p25.3) and *MYC* (8q24.21) are amplified in the CCA-Like, Blast-Like, and HCC tumors (**Figure 2.7.5c-d**). The CCA-Like and Blast-Like classes displayed numerous differences in copy number landscape frequency compared to CCA and HCC (**Figure 2.7.5a**). Blast-Like tumors had increased genomic instability, which resulted in a more distinct copy number landscapes, 1143 segments were significantly different as compared to HCC and 3489 segments as compared to CCA with 2690 segments shared as significantly different compared to both CCA and HCC (**Supplemental Figure 2.8.7b**). The gene expression was compared to the GISTIC copy number values for *BAP1*, *FOXC1* and *MYC*. Decreased *BAP1* gene expression was correlated to a decreased to the copy number status for all classes (across all samples, p-value = 2.2e-16, CCA, CCA-Like, Blast-Like and HCC $p<0.001$) (**Supplemental Figure 2.8.7c**). Although *MYC* was amplified in CCA-Like, Blast-Like, and HCC, only in the Blast-Like and HCC classes was gene expression and copy number correlated (Blast-Like and HCC p-value < 1.0e-4), whereas gene expression and copy number were not correlated in any class for *FOXC1* (**Supplemental Figure 2.8.7d-e).**

2.4.7 CCA-Like and Blast-Like have decreased progression-free survival

Progression-free and overall survival censored at five years was compared across classes (**Figure 2.7.6a-b**) (**Supplemental Table 2.8.3**). Using a Cox proportional hazards model, Blast-Like tumors had significantly worse progression-free (Hazard Ratio [HR] 1.95, p<0.001) and overall (HR 3.72, p<0.001) survival comparted to HCC. CCA-Like tumors had worse progression-free survival compared to HCC (HR 1.68, p-value =0.04), but not overall survival. We also looked at models including the clinical factors stage and grade. In a univariate model, only stage was associated with outcomes. When we added stage to the model with our subclasses, only the Blast-like class retained significance for worse progression-free and overall survival compared to the referent class HCC. AFP protein expression has additionally been shown to be a prognostic marker, as such, we performed a multivariate analysis; when combining AFP with subclasses, Blast-Like and CCA-Like were still significant predictors of worse progression free (HR=1.9, p=0.002 and HR=1.6, p=0.03, respectively) and Blast-Like was significant predictor of worse overall survival (HR=3.5, p=8e-10). We generated a gene expression classifier for our classes based on TCGA data and applied it to Roessler et. al. HCC cohort as a validation dataset (GSE14520)[149]. The five-year survival within the validation cohort displayed similar trends to TCGA. The Blast-Like class trended towards a worse relapse-free survival (HR 1.4, p-value=0.055) and had significantly worse overall survival with and without adjusting for stage (HR 1.9, p-value=0.006, HR 2.0, p-value=0.002, respectively) (**Figure 6c-d**, **Supplemental Table 2.8.3**).

We next applied the Hoshida et. al. and Woo et. al. signatures to our TCGA dataset using consensus clustering. Hoshida identified 3 subclasses (S1, S2 and S3) that were correlated to clinical and molecular features[88]. Woo et. al. identified a cholangio-like group of tumors (CLCHCC), and then further divided the subtypes with respect to stemness[129]. With both the Hoshida and Woo subtyping strategies, the CCA-Like and Blast-Like samples were grouped

together within the S2 and CLCHCC classes respectively (**Supplemental Figure 2.8.8a-c**). The stem cell signature from Ben-Porath et. al.[135] was applied accordingly to the Woo et. al. dataset and the CLCHCC class was further divided into stem cell signature positive or negative (**Supplemental Figure 2.8.8c**). Overall, all groups identify distinct subtypes with similar, but not completely overlapping features. Our scheme helps solidify the classifications by anchoring with true CCA and incorporating microdissected and single cell data.

Progression-free and overall survival curves were generated using TCGA data for both our classes as well as Hoshida and Woo classifications (**Supplemental Figure 2.8.8d-i**). For progression-free survival, Blast-like and CCA-Like had the shortest time to progression with a median time of 301 days and 355 days, respectively, as compared to HCC (879 days) (p=1.2e-9, p=0.01) (**Supplemental Figure 2.8.8d**). Within the Hoshida subtypes, no significant difference in survival was observed (**Supplemental Figure 2.8.8e**). The median time to progression for the poorest outcome CLCHCC (ES) class was 355 days (p=2.0e-4) as compared to CLCHCC (neg ES) and HCC (neg ES), which had median time to progression >700 days (**Supplemental Figure 2.8.8g**). Our Blast-like classification identified a subset of patients with poorer outcomes.

2.5 Discussion

In this study, we used a combined TCGA CCA/HCC dataset to characterize HCC samples based on their similarity to CCA and the precursor hepatoblast cell type. Previous studies which have noted similar transcriptional classes [127-129]. Here, we have expanded on these classifications and used a multi-omic approach as well as using external data to characterize rare molecular subtypes in a fully integrated manner; the result of which is the identification of three distinct classes of HCC tumors: CCA-Like, Blast-Like and HCC.

Previous work had identified *IDH1* mutant and CCA-like subclasses of tumor similar to the aforementioned CCA-Like. Our work is a natural extension of the groundwork laid in these papers. Woo et al identified a cholangiocarcinoma-like group with varying levels of embryonic stem (ES) cell marker expression[129]. Whereas the expression of ES signatures led them to conclude that these tumor were derived from a bi-potent progenitor cells, hepatoblast; our current study, with the addition of copy number data, demonstrates that the CCA-like group have a copy number landscape that more closely resembles HCC. This in combination with expression of liver specific genes, albeit at reduced levels, is an indication that the CCA-like class is more likely to be derived from hepatocytes that were transformed and underwent dedifferentiation and initiated a transdifferentiation transcriptional program in response to specific mutations (e.g., BAP1, IDH1/2). The Cancer Genome Project previously described IDH1/2 mutant HCC and presented evidence that these tumors overlapped CCA via a TumorMap visualization, that incorporated DNA, DNA methylation and expression features; however, this group was restricted to small subset with IDH1/2 mutation[127]. Our work expanded this group to include a set of transcriptionally similar tumors that includes samples with a high frequency of *IDH1*/2 mutations and alterations in *BAP1*, as well as linking this subclass to the induction of a transdifferentiation program.

A majority of the tumors originating from liver displayed classical HCC features including expression of *ALB* and *HNF4A* genes as well as mutations in *CTNNB1* (29%). However, Blast-like tumors exhibited more frequent mutations in *TP53* (58%) and elevated *AFP* expression, a hepatoblast marker gene. It has been previously reported that *TP53* mutations, specifically R249S mutations, are most commonly observed in east Asian populations and are associated with aflatoxin exposure [166]. Corroborating this R249S/Aflatoxin relationship, we saw that tumors with the R249S mutations also had high similarity to a previously described Aflatoxin mutational signature (**Supplemental Figure 2.8.5d**). Furthermore, the Blast-like class was enriched with

patients of Asian ancestry (**Supplemental Table 2.8.1**) as well as having a significantly higher rate of HBV infection (**Supplemental Table 2.8.1**) (**Figure 2.7.2a**, **Supplemental Figure 2.8.3a-b**). The association between Asian ancestry and HBV+ HCC is expected, as the historic prevalence of HBV infection in East Asia is ~7.3% as compared to the North America at 0.3% [167]. Additionally, many of the samples of Asian ancestry within the TCGA come from Asian tissue source site. Current data has linked chronic HBV infection to immune induced liver injury[168-170], this injury can result in dedifferentiation of hepatocyte[171] and in turn, based on our data, lead to a more hepatoblast-like disease in east Asian populations.

CCA-Like lacks prototypical mutations and risk factors associated with HCC. Hirsch et. al. previously described a BAP1 mutant class of tumors that lacked *CTNNB1* mutation and canonical risk factor but has similarity to fibrolamellar tumors[172]. Our work builds on this group as CCA-Like tumors are enriched for *IDH1/2* and *BAP1* mutations and have transcriptional patterns similar to CCA. By using transcriptional patterns to define the CCA-Like class, we found that in addition to *BAP1* mutations *BAP1* copy number loss was another frequent mechanism for decreasing *BAP1* mRNA expression levels (**Supplemental Figure 2.8.7c**).

By performing mutational signature profiling we observed the CCA-Like had a more similar global mutational signature profile to CCA than HCC. However, CCA-Like still showed underlying exposure-based signatures though their overall contribution to the mutation burden was reduced compared to HCC (**Figure 2.7.3e-f**). A hepatocellular cell-of-origin is reinforced by the observation that the CCA-Like and HCC tumors have similar expression levels of *ALB* and *AFP*. The CCA-Like tumors were also classified as hepatocellular carcinomas by TCGA's expert pathology re-review. Woo et. al. has identified a similar subclass (Cholangiocarcinoma-Like), but as previously mentioned these authors attribute this to transformation of a progenitor hepatocellular cell vs a hepatocyte [129]. However, the CCA-Like tumors have increased *ALB* expression and liver specific gene and miRNA expression signatures when compared to CCA,

suggesting still a strong linkage to hepatocytes. More recently Wardell et. al. suggested that genomic features of intrahepatic CCA tumor suggest a hepatocyte cell of origin[173]. While our work is an agreement with that notion, we extend their finding through transcriptional analysis proposed transdifferentiation pathways. We show that CCA-Like have upregulation of NOTCH, WNT and TGFβ pathways as compared to Blast-like and HCC, all of which are known to be associated with transdifferentiation. A murine study by Sekiya and Suzuki demonstrated that NOTCH signaling in hepatocytes can induce the conversion of hepatocytes to biliary cells leading to the development of cholangiocarcinoma [159]. Additionally, in NOTCH deficient mice, expression of *TGFB* allows the generation of the biliary tree from hepatocytes[158].

This notion of transdifferentiation is bolstered by the finding that the copy number landscape of CCA-Like is more similar to HCC than CCA (**Figure 2.7.5a**). Copy number alterations are thought to be early events in tumor development; this data corroborates the mRNA data by suggesting CCA-Like is derived from a hepatocyte rather than a bipotent progenitor cell. Additionally, the CCA specific DNA alterations, *IDH1/2* and *BAP1*, could be driving this transdifferentiation process in the absence of classical HCC mutations such as *CTNNB1* and *TP53*. Artegiani et. al. recently reported that BAP1-/- organoids upregulated EPCAM while downregulating liver specific genes, consistent with our findings (**Supplemental Figure 2.8.4a**) [174].

Prior groups have laid a strong basis for subtype classification including the identification of subtypes similar to the CCA-like[88,128,129,172] and Blast-like[88-90]. We have added to this rich body of work with a direct comparison to true CCA and incorporated mutational signature analysis and copy number data to further describe the underlying biology and potential cell of origin for these classes. Our findings link a specific class of HCC tumors with transdifferentiation; however, further work will need to be done to validate this mechanistically to identify the exact

genomic alterations and signaling pathway changes that are necessary and sufficient to drive the CCA-Like tumor type.

One limitation of our study was the partial availability of TCGA's non-required data elements including: serum markers, family histories, consistently annotated risk factors and long-term follow up. More in-depth and standardized annotation for these clinical data elements will be important to better understand associations between the molecular data and etiologic risk factors.

Through the integration of multiple data types, we were able to expand on prior work, which identified a subset of HCC tumors that resembled CCA. We identified three distinct classes of hepatocellular carcinoma, in which the CCA-Like class may be derived through initiation of a transdifferentiation process, rather than transformation of progenitor cell.

2.6 Acknowledgements

We thank the members of the K.A.H. and D.N.H laboratories for useful discussions. Aspect of **Figure 2.7.1** were created with BioRender.com and http://repo.mouseimaging.ca/repo/4D_embryo_atlases_M_Wong/. This work was supported by National Institutes of Health Grant 5 U24 CA210988 (to K.A.H.) and National Institutes of Health (Program in Translational Medicine T32 GM122741 to UNC/M.A.S).

2.7 Figures and Tables

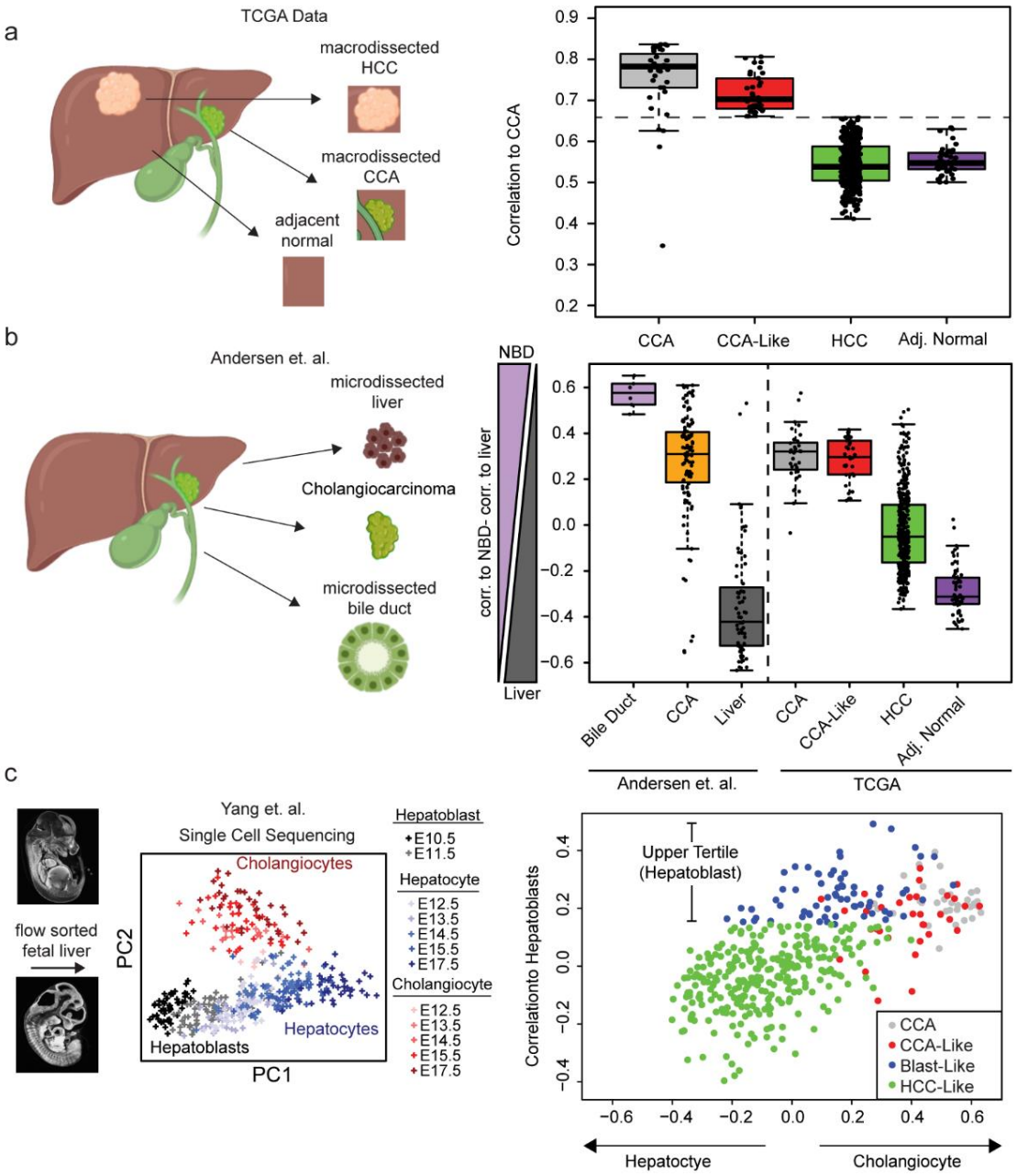


Figure 2.7.1. Molecular classification of hepatocellular carcinoma

(a) The Spearman correlation to the median expression of CCA tumors (n=36) was calculated for each TCGA CCA/HCC tumor (n=410). HCC samples within ±1 standard deviation of the mean CCA Spearman correlation (dashed line) were defined at CCA-Like. **(b)** The CCA/HCC

dataset was correlated to microdissected normal bile duct (NBD) (n=6) or normal liver (n=59) from Andersen et. al. An NBD vs. Liver score was calculated by subtracting the correlation to normal liver from the correlation to normal bile duct [150]. Boxes represent the IQR with median represented by the bolded bar. Error bars represent Q1/Q3 ± 1.5*IQR. **(c)** Single cell RNA seq data from Yang et. al. was used to correlate the CCA/HCC samples to either hepatoblasts (E10.5, n=54), hepatocytes (E17.5, n=34) and cholangiocytes (E17.5, n=34). PCA was performed on Yang et. al. to visual variance across the samples then the correlation was calculated between the median expression of day E10.5 and E17.5 (hepatocytes and cholangiocytes) samples to the CCA/HCC dataset. HCC samples in the upper tertile of correlation to hepatoblasts and not prior classified as CCA-Like, were defined as Blast-Like (n=66). Murine embryo images were obtained from http://repo.mouseimaging.ca/repo/4D_embryo_atlases_M_Wong/.

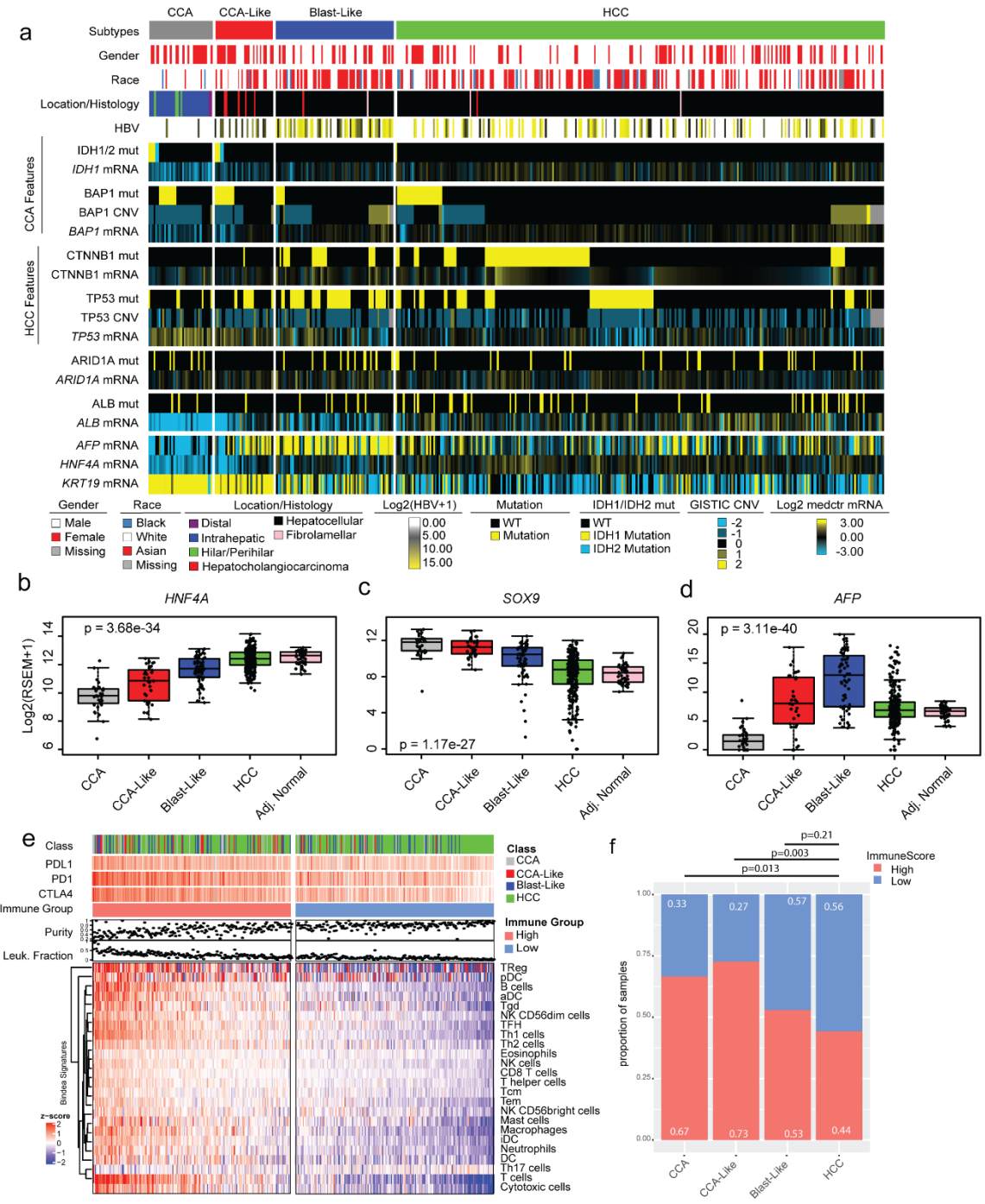


Figure 2.7.2. CCA-Like tumors are molecularly similar to CCA

(a) Samples are ordered by subclass and sorted by key genomic alterations. Gene expression data is log2 transformed and median centered. Mutations (mut) are indicated by yellow/blue while wildtype (WT) are indicated by black. GISTIC thresholded values were used for copy

number variation (CNV). **(b-d)** Gene expression for hepatocyte markers (*HNF4A and AFP*) and cholangiocyte marker (*SOX9*) are shown per CCA/HCC group. Gene expression values represent the log2 transformed RSEM+1 value. One-way ANOVA p-value is displayed. Boxes represent the IQR with median represented by the bolded bar. Error bars represent Q1/Q3 ± 1.5*IQR. **(e)** The median expression across the Bindea immune signatures were calculated and clustered by signatures. Samples were sorted by decreasing median expression across all the signature and divided into high and low expression groups. PDL1 [range= 0, 8.5], PD1 [0, 11.1] and CTLA4 [0, 9.9] are represented as Log2(RSEM+1 values). **(f)** Stacked bar plots represent the proportion of high and low immune expression group across the tumor classes. Fisher exact test p-values are shown for each comparison to HCC.

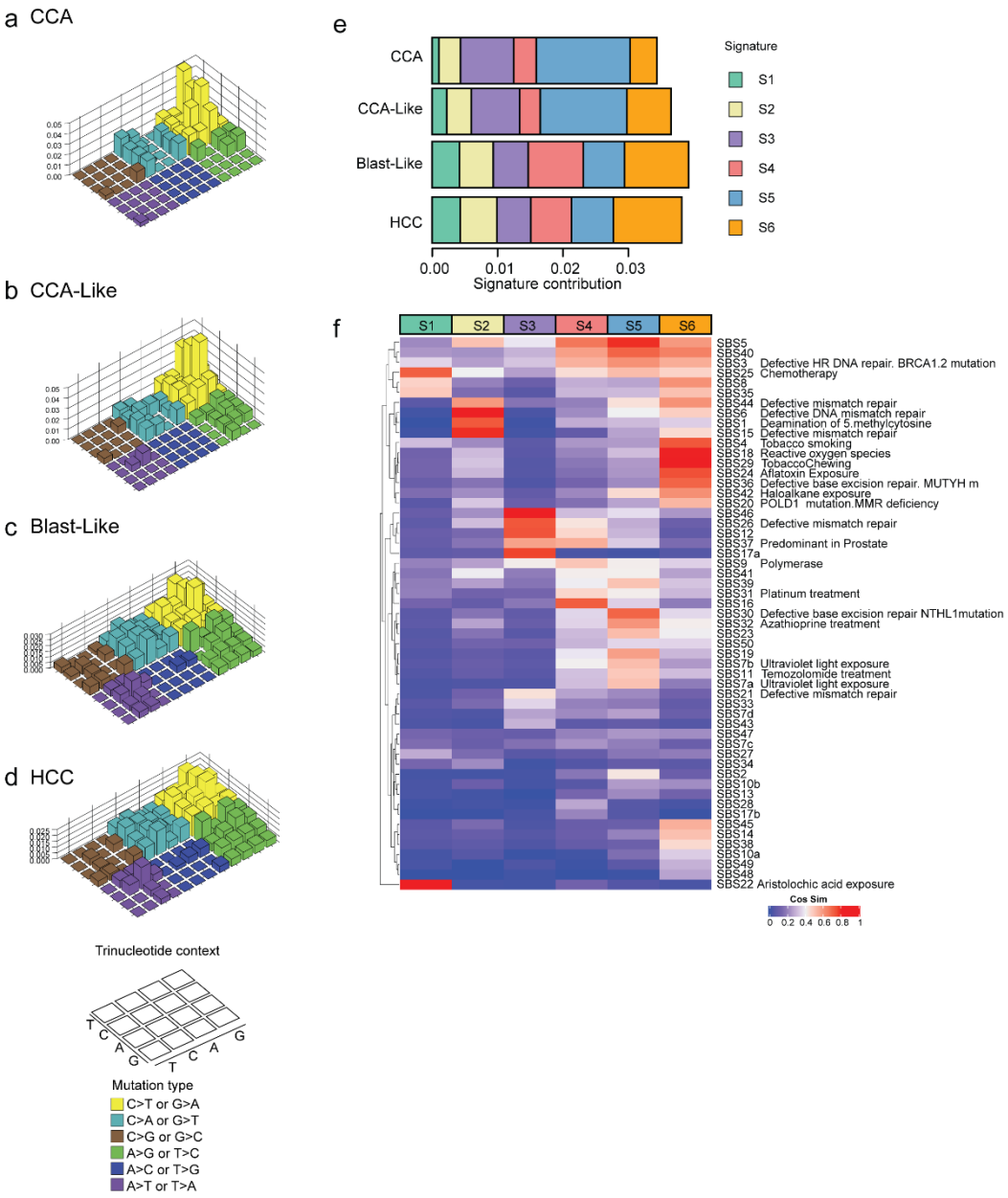


Figure 2.7.3. Mutational signatures of CCA and HCC reveal cross class similarity

(a-d) The median frequency of each single nucleotide variant (SNV) per class was calculated and plotted by preceding and succeeding base in a Lego plot (key on bottom left). Base substitutions are divided into six categories to represent the six possible base changes. Substitutions are further divided by the 16 possible flanking nucleotides surrounding the mutated base as listed in the trinucleotide context legend. **(e)** The R package SomaticSignatures was used to identify *de novo* mutational signatures. Signature contribution

across samples is aggregated by class and the median contribution of each signature to the tumor classes is shown. **(f)** Cosine similarity between COSMIC v3 mutational signatures and each of the *de novo* signatures were computed. The COSMIC signatures are clustered by cosine similarity and ordered by signature class (S1-S6). Color key indicates the degree of similarity.

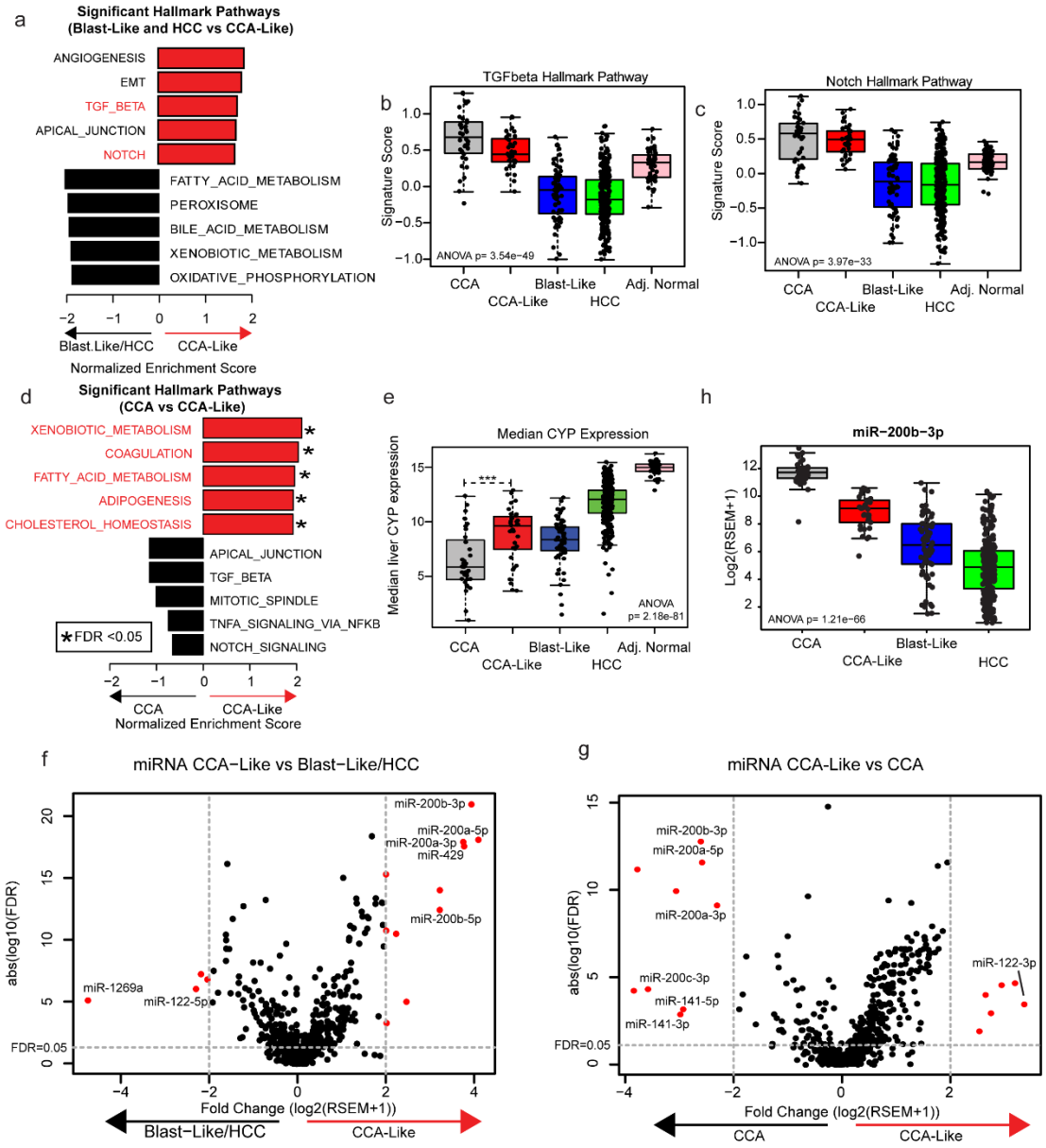


Figure 2.7.4. CCA-Like tumors display features of transdifferentiation

(a) Gene Set Enrichment Analysis was performed 1 vs Rest (CCA-Like vs. Blast-Like and HCC) using the Hallmark gene sets. Seven gene sets were enriched in CCA-Like versus Blast-Like and HCC, and gene sets associated with transdifferentiation are noted in red. **(b-c)** Signature scores associated with transdifferentiation pathways, TGFβ and Notch are plotted by tumor classification. One-way ANOVA p-values are shown. **(d)** Gene Set Enrichment Analysis

comparing CCA-Like vs. CCA using the hallmark gene sets. Liver specific gene sets are noted in red. **(e)** The median expression of liver specific cytochrome P450 genes are plotted by subtype, *** indicates p<0.001 for two sample t-test between CCA and CCA-Like. **(f-g)** Volcano plots for fold change vs FDR; are plotted for CCA-Like vs. Blast-Like and HCC and CCA-Like vs. CCA respectively; genes with fold change > 2 and FDR < 0.05 are indicated in red. **(h)** Per class expression of miR-200b-3p, a representative family member of the miR-200 family is plotted, p-value represents ANOVA. Boxes represent the IQR with median represented by the bolded bar. Error bars represent Q1/Q3 ± 1.5*IQR.

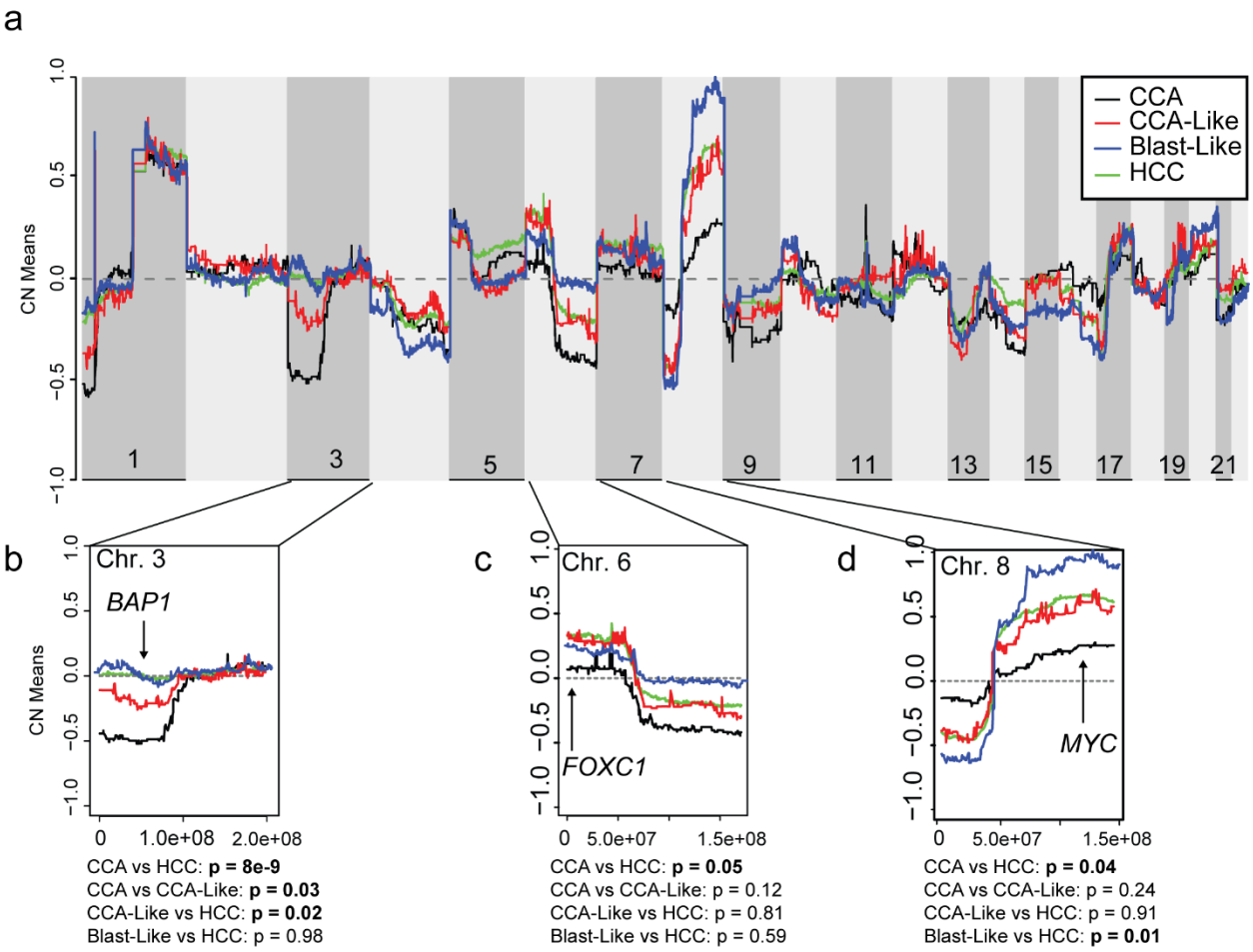


Figure 2.7.5. Copy number landscapes show tissue specific specificity

(a) Genome-wide copy number values are plotted for CCA (black), CCA-Like (red), Blast-Like (blue), and HCC (green) using the mean quantitative gene level copy measurements from GISTIC. **(b-d)** Expanded view of three chromosomes containing regions significantly differed by pairwise two sample t-tests (p-value <0.05) between CCA and HCC. Potential target genes in significant segments are noted.

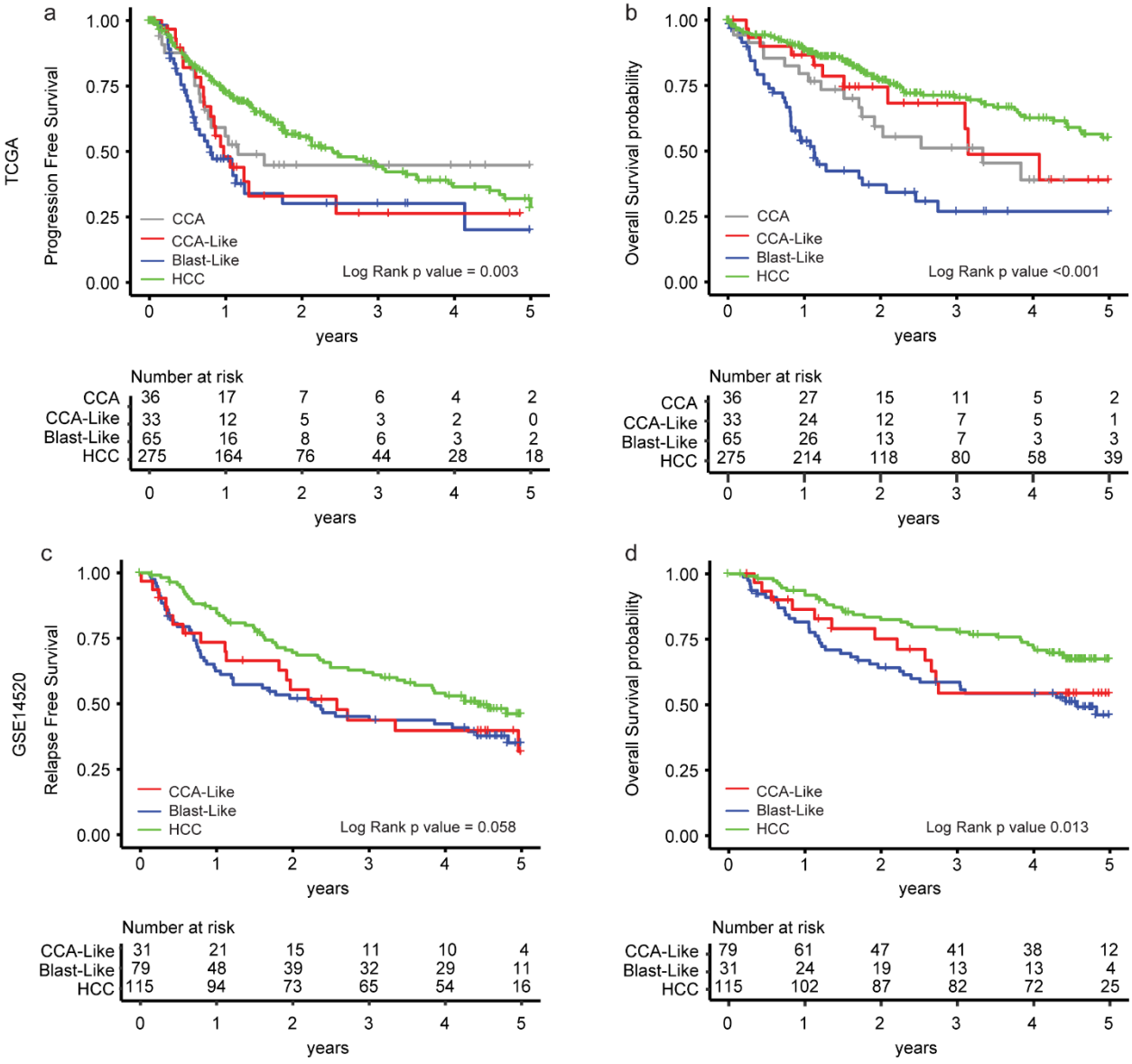
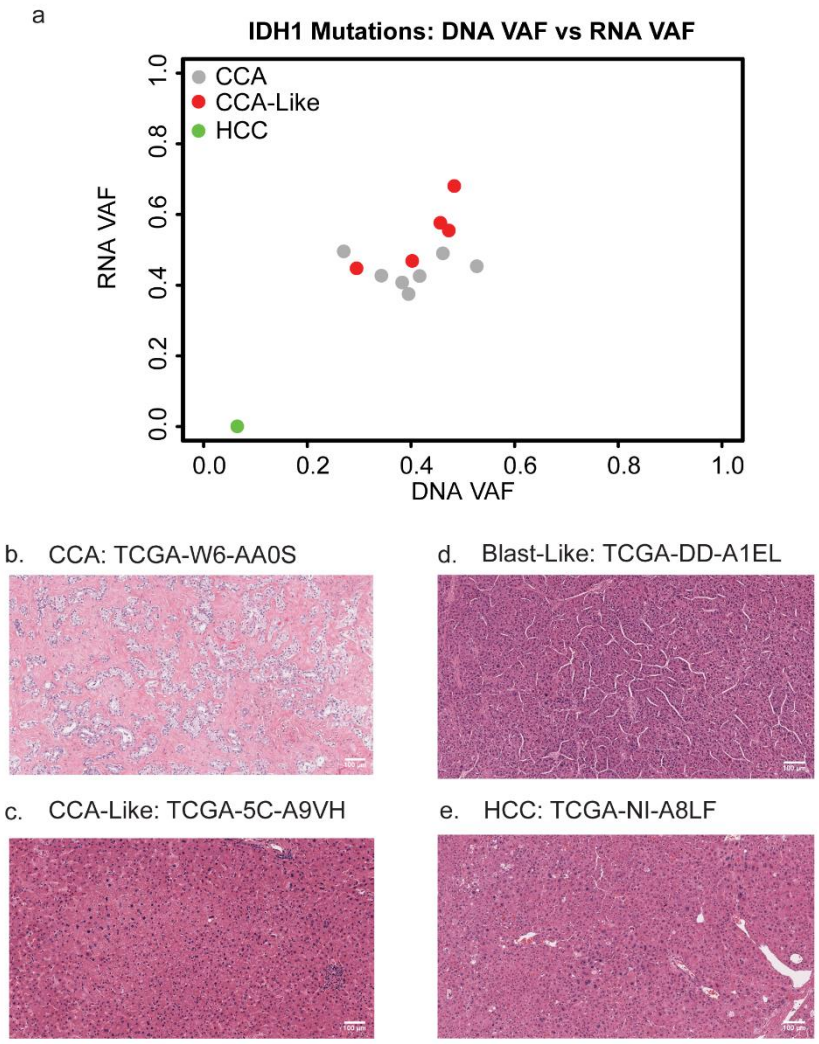


Figure 2.7.6. Blast-Like tumors have worse progression-free and overall survival

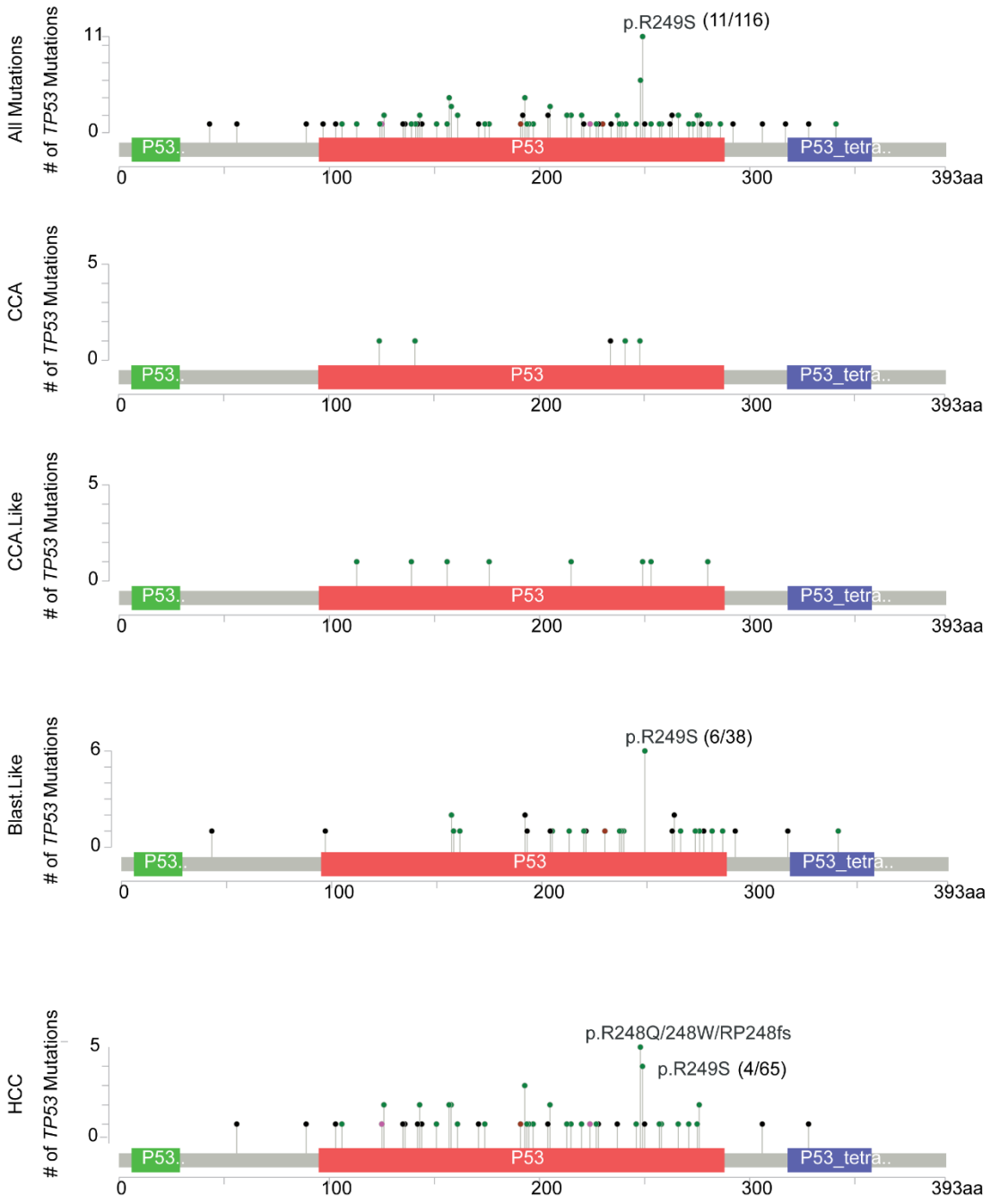
Kaplan-Meier curves of TCGA CCA/HCC data for **(a)** progression-free and **(b)** overall survival. Tumor classifications were predicted on to GSE14520 using ClaNC for **(c)** relapse-free and **(d)** overall survival[149]. All survival data was censored at 5 year and log-rank p-value was calculated.

2.8 Supplemental Figures and Tables



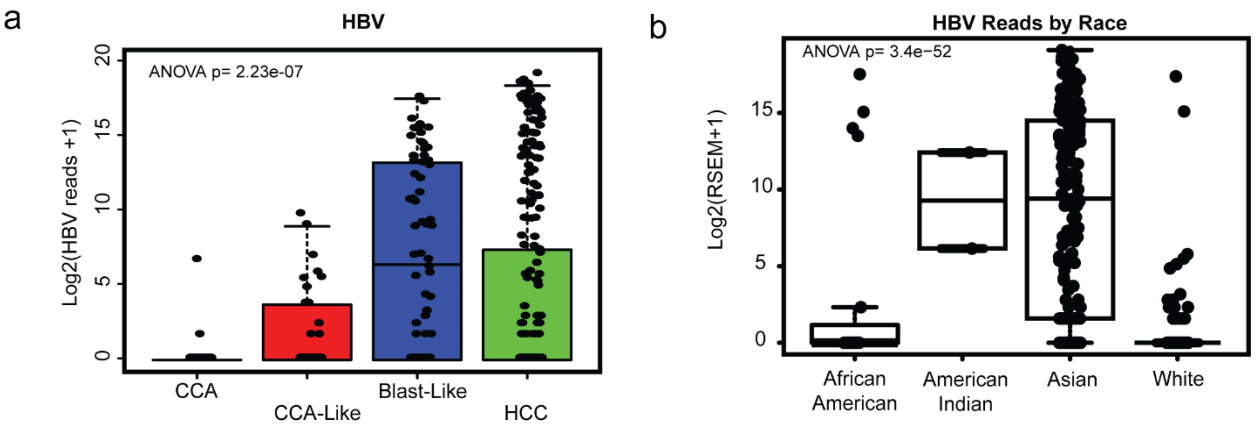
Supplemental Figure 2.8.1. CCA-Like share CCA genomic alterations but has HCC histological features

(a) Variant allele frequencies for IDH1 were calculated from both the DNA and RNA sequencing reads. Tumor classes are indicated by color (grey=CCA, red=CCA-Like, green=HCC). **(b)** Representative H&E slides (20x) of CCA **(c)** CCA-Like **(d)** Blast-Like **(e)** HCC. (https://cancer.digitalslidearchive.org/).



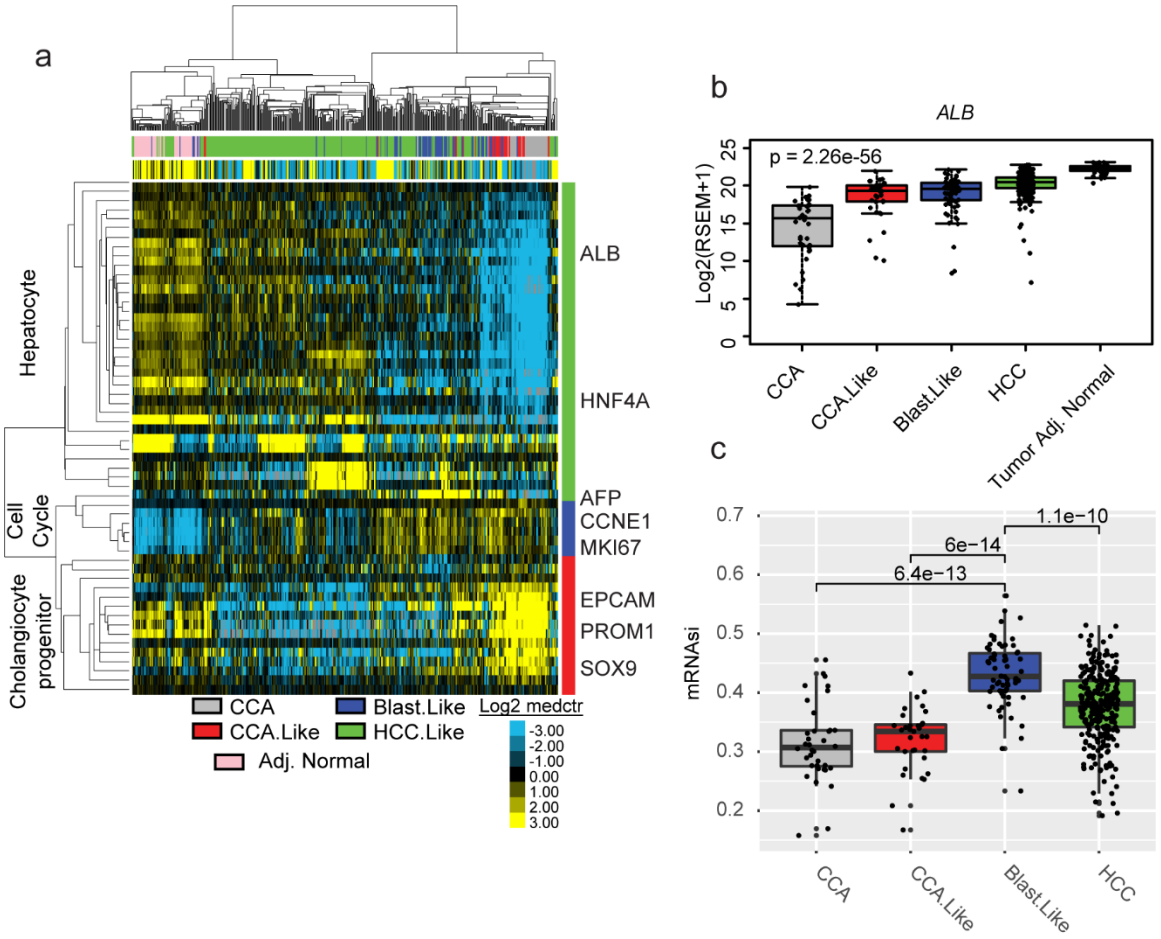
Supplemental Figure 2.8.2. Lollipop plots for *TP53* mutations

Lollipop plots for *TP53* mutations are shown for the entire cohort and by tumor class. R249S mutation ratios are shown as (number of R249S mutations)/ (total number of TP53 mutations) per group.



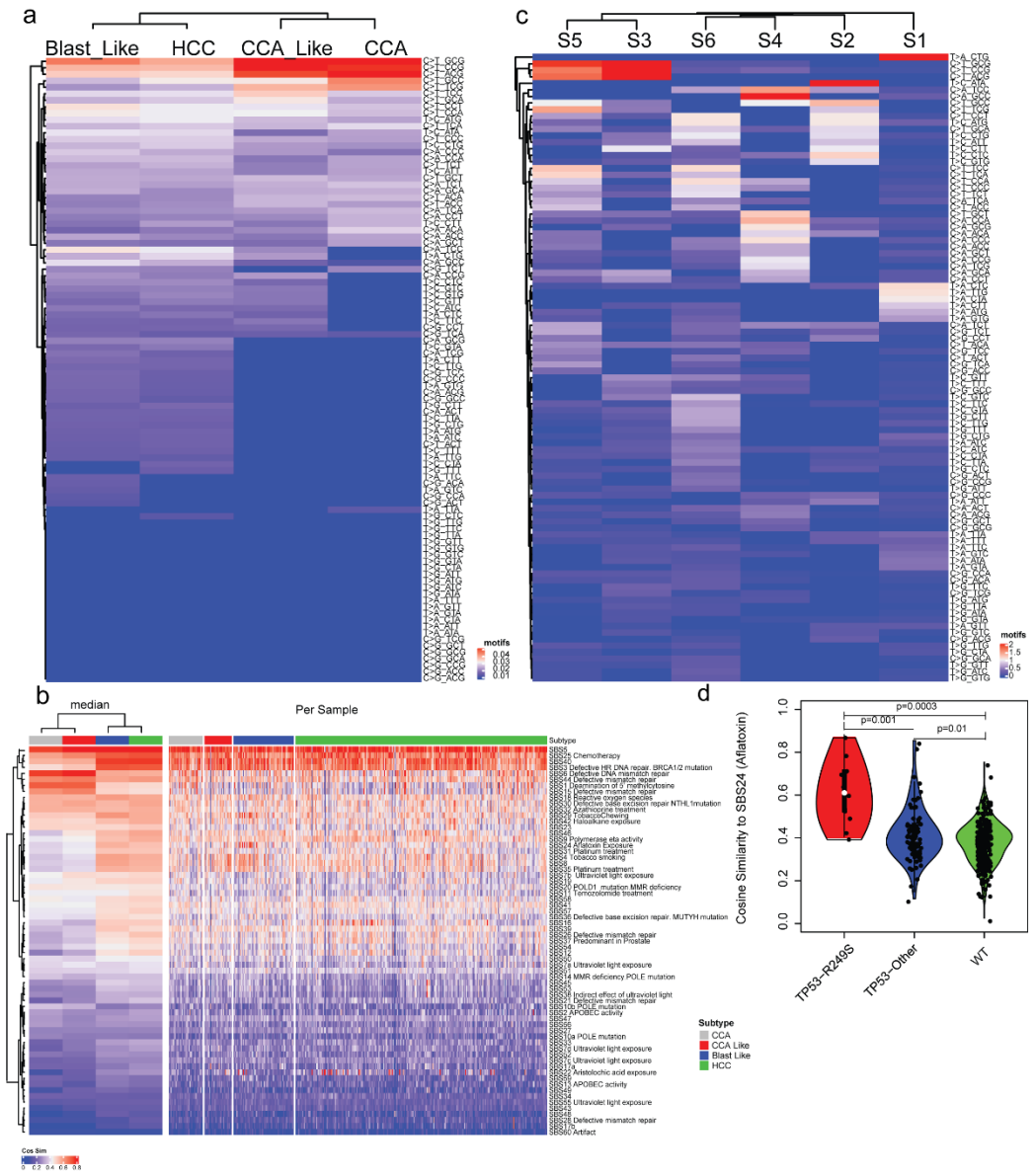
Supplemental Figure 2.8.3. High HBV reads in Blast-like samples associates with Asian ancestry

(a) Reads aligning to HBV are plotted by tumor class and **(b)** reported race



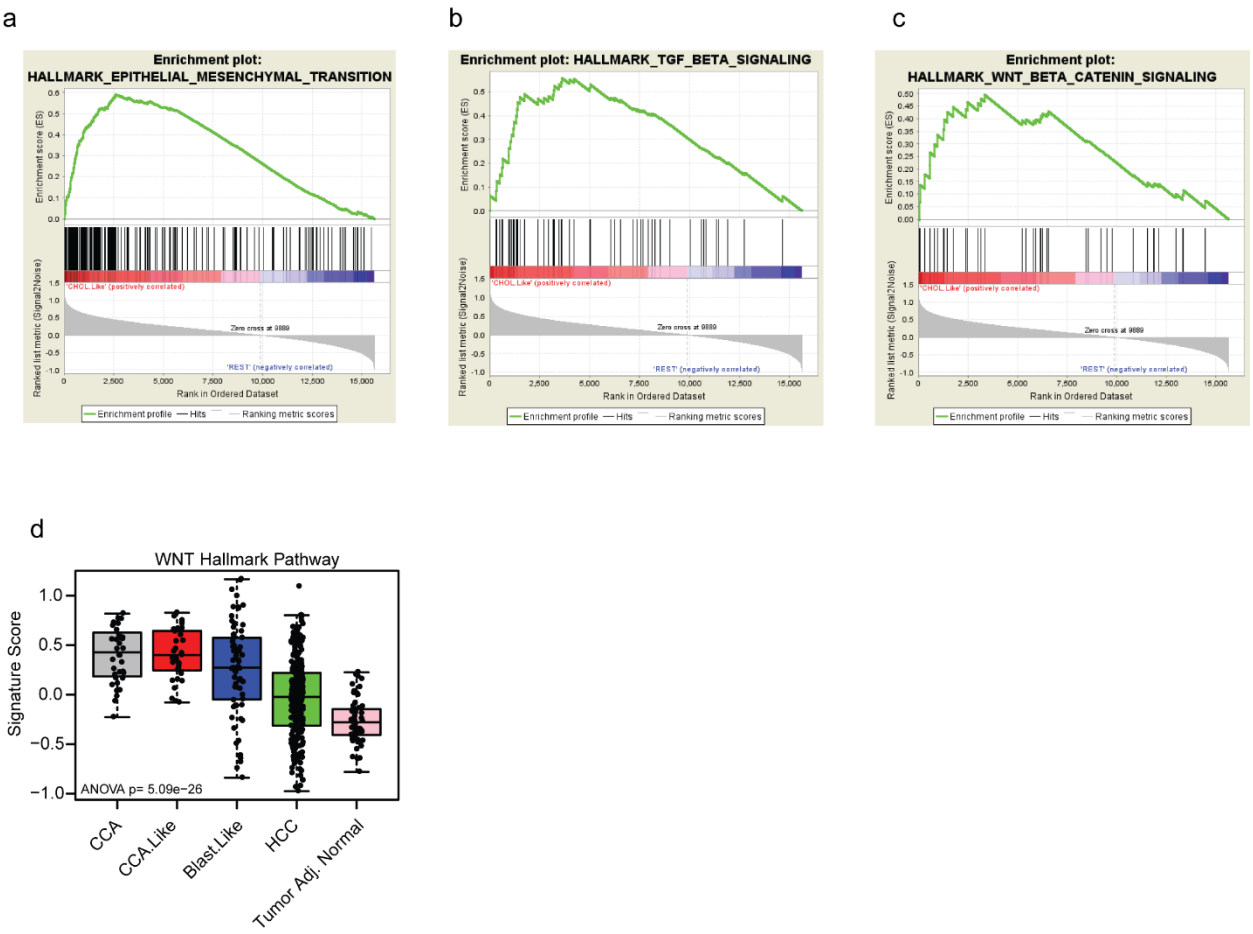
Supplemental Figure 2.8.4. CCA-Like shared gene-expression features with CCA tumors

(a) CCA/HCC tumors and adjacent normal tissues clustered with cell type specific markers of hepatocytes and cholangiocytes cells and cell cycle markers from Hu et. al.[133]. Gene expression data was log2 transformed and median centered across the CCA/HCC cohort. Annotation bar represents sample subclasses. **(b)** Gene expression for *ALB* is shown per CCA/HCC group (ANOVA p=2.26e-56). Gene expression values represent the log2 transformed RSEM+1 value. **(c)** The expression based stemness index from Malta et. al.[153] was plotted by class. p-values represents t-test.



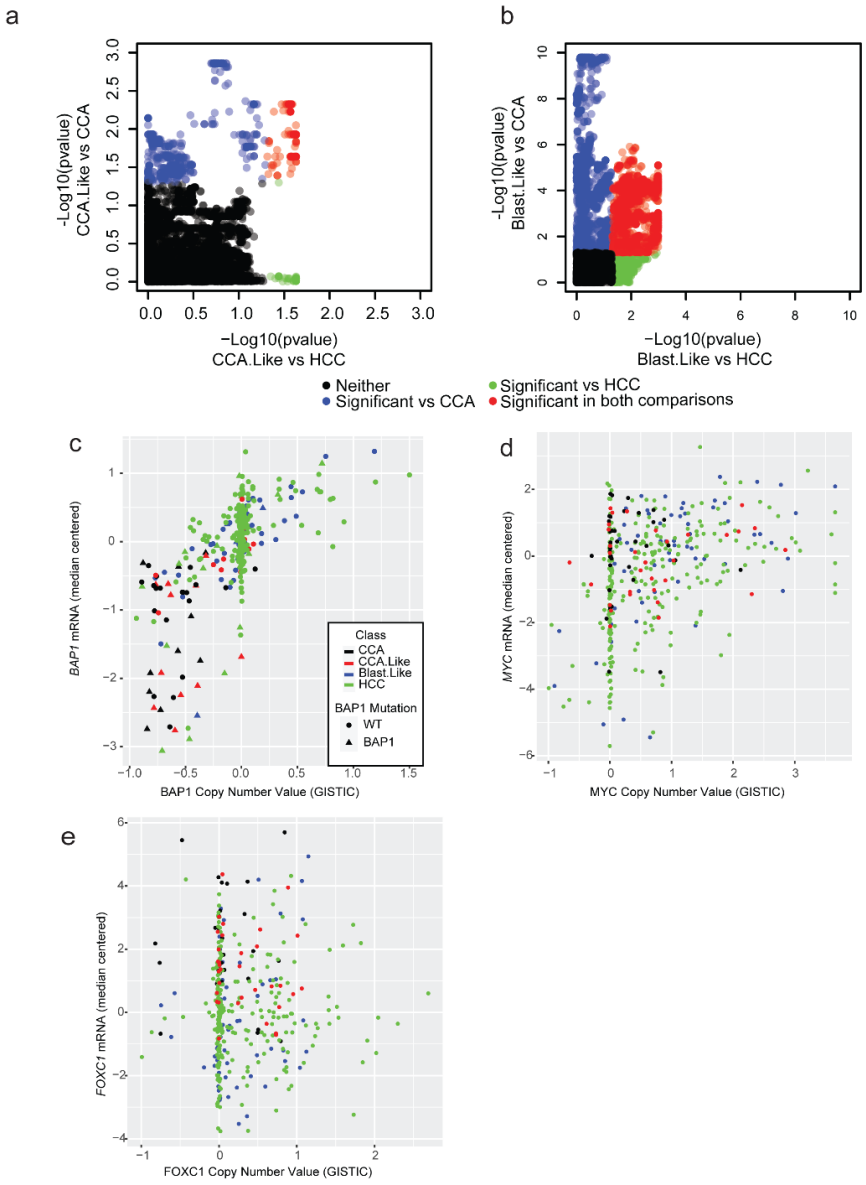
Supplemental Figure 2.8.5. CCA-like and CCA classes share mutational patterns

The median motif frequency was hierarchically clustered by motifs for (a) classes. (b) COSMIC signatures were clustered by median class cosine similarity. The per sample CS was then similarity ordered and sorted by class. (c) The median motif frequency was hierarchically clustered by motifs for signatures. (d) Cosine similarity for SB24 (aflatoxin exposure) was plotted by *TP53* mutation type (TP53 R249S, Others or WT) and pairwise t-test were performed to determine significance.



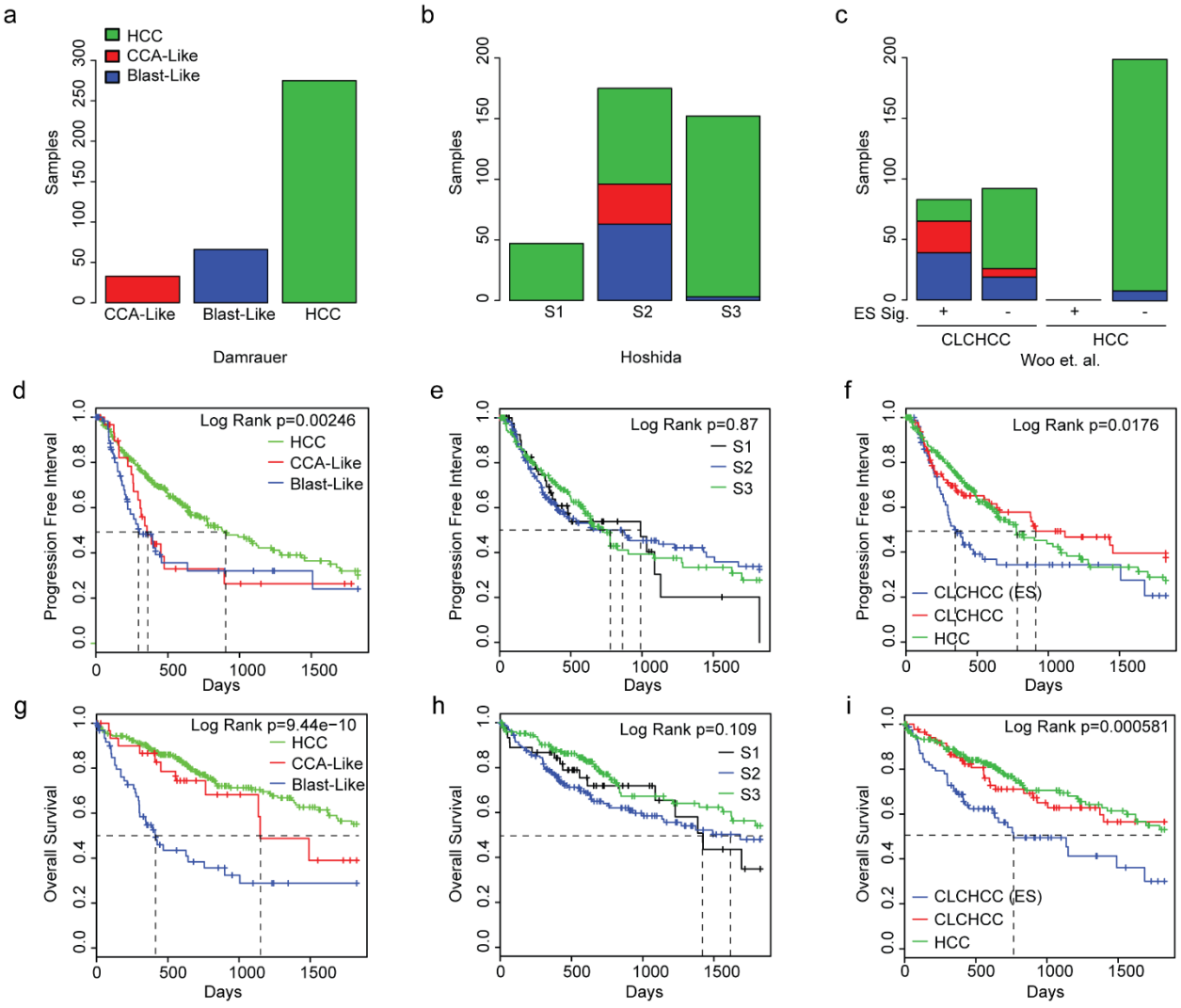
Supplemental Figure 2.8.6. CCA-like display elevated expression of transdifferentiation-associated pathways compared to HCC classes

(a-c) GSEA was performed comparing CCA-Like vs. Blast-Like/HCC. Plots of enrichment scores are shown for pathways relating to transdifferentiation. **(d)** Signature score associated with WNT pathway are plotted by class.



Supplemental Figure 2.8.7. CCA-like copy number landscape resembles HCC, while Blast-like has greater genomic instability compared to other HCC classes

(a) Pairwise two sample t-tests were performed by segment and p-values were corrected for multiple comparisons. –log10 (adjusted p-value) was plotted for CCA-Like vs CCA and HCC and **(b)** Blast-Like vs. CCA and HCC. Log2 median centered gene expression was plotted against the GISTIC intensity values for **(c)** *BAP1*, **(d)** *MYC* and **(e)** *FOXC1*. For *BAP1* mutations noted by symbols.



Supplemental Figure 2.8.8. Blast-like classification schema identified a subset of patients with worse outcomes

(a-c) Hoshida et. al. and Woo et. al. subtype classifications were applied to TCGA cohorts. **(d-f)** Progression-free and **(g-i)** overall survival was plotted for the given tumor classifications. Dashed lines represent 50% survival.

Supplemental Table 2.8.1. Clinical Characteristics of TCGA CCA and HCC datasets

	CCA	CCA-like	Blast-like	HCC
No. of Cases				
	36	33	66	272*
Gender				
Female	20 (56%)	20 (61%)[b,c]	21 (32%)	80 (29%)
Male	16 (44%)	13 (39%)	45 (68%)	192 (71%)
Race				
African American	2 (6%)	1 (3%)	3 (5%)	13 (5%)
Asian	3 (8%)	12 (36%)	40 (60%)[e]	106 (39%)
White	31 (86%)	20 (61%)	22 (33%)	142 (52%)
Other/NA	0 (0%)	0 (0%)	1 (2%)	11 (4%)
Stage				
I	19 (53%)[b]	12 (36%)	18 (27%)	141 (52%)[a,b]
II	9 (25%)	7 (21%)	22 (33%)	57 (21%)
III	1 (3%)	10 (30%)	24 (36%)	51 (19%)
IV	7 (19%)	0 (0%)	1 (2%)	4 (1%)
NA	0 (0%)	4 (12%)	1 (2%)	19 (7%)
Grade				
G1	1 (3%)	1 (3%)	5 (8%)	49 (18%)[e]
G2	15 (42%)	14 (42%)	20 (30%)[c]	143 (53%)
G3	18 (50%)	14 (42%)	36 (54%)[c]	72 (26%)
G4	2 (5%)	3 (9%)[c]	5 (8%)[c]	4 (1%)
NA	0 (0%)	1 (3%)	0 (0%)	4 (1%)
ECOG Score				
0	20 (56%)	14 (42%)	17 (26%)[c,d]	131 (48%)
1	9 (25%)	5 (15%)	10 (15%)	69 (25%)
2	0 (0%)[b]	1 (3%)	11 (17%)[c,d]	14 (5%)
3	1 (3%)	1 (3%)	6 (9%)[d]	5 (2%)
4	0 (0%)	0 (0%)	3 (5%) [d]	0 (0%)
NA	6 (17%)	12 (36%)	19 (29%)	53 (19%)
Histology				
Fibrolamellar	0 (0%)	0 (0%)	1 (2%)	2 (1%)
Hepatocellular	0 (0%)	28 (85%)	64 (97%)	269 (99%)
Hepatocholangiocarcinoma	0 (0%)[a]	5 (15%)[e]	1 (2%)	1 (0%)
HBV infection [f]				
Positive	2 (6%)	12 (36%)	44 (67%)[e]	97 (36%)

Negative	34 (94%)	21 (64%)	22 (33%)	174 (64%)
NA	0 (0%)	0 (0%)	0 (0%)	4 (1%)
Cirrhosis/Fibrosis				
No Fibrosis	16 (44%)[e]	5 (15%)	7 (11%)	62 (23%)
Fibrosis	11 (31%)	7 (21%)	5 (8%)[d]	47 (17%)
Cirrhosis	0 (0%)[e]	5 (15%)	10 (15%)	64 (24%) [d]
NA	9 (25%)	16 (48%)	44 (67%)	99 (36%)

[a] Significant as compared to CAA.Like ($p<0.05$)

[b] Significant as compared to Blast.Like ($p<0.05$)

[c] Significant as compared to HCC ($p<0.05$)

[d] Significant as compared to CAA ($p<0.05$)

[e] Significant as compared to all other groups ($p<0.05$)

[f] HBV infection determined by >5 HBV reads from RNA-seq

*272/275 samples had unique clinical annotation

Supplemental Table 2.8.2. Molecular Characteristics of TCGA CCA and HCC datasets

	CCA	CCA-like	Blast-like	HCC
No. of Cases				
	36	33	66	275
IDH				
IDH1/IDH2 mutation	6	5[b,c]	0	1
IDH1 mRNA (median)[f]	11.63[e]	12.43[e]	12.84	13.29
BAP1				
BAP1 mutation	10 (27%)[b,c]	11 (33%)[b,c]	5 (8%)	25 (9%)
BAP1 copy number loss	29 (80%)[e]	15 (45%)[c,d]	17 (26%)	33 (12%)
BAP1 mRNA (median)[f]	10.3	10.73	11.07	11.12
CTNNB1				
CTNNB1 mutation	0 (0%)	1 (3%)	18 (27%)[a,d]	79 (29%)[a,d]
CTNNB1 mRNA (median)[f]	11.84[b,c]	12.43[b,c]	12.6	12.44
TP53				
TP53 mutation	5 (14%)	8 (27%)	38 (58%)[e]	65 (24%)
TP53 copy number loss	13 (36%)[e]	23 (70%)	45 (68%)	152 (55%)
TP53 mRNA (median)[f]	10.83[e]	10.28[e]	9.86	9.55
ARID1A				
ARID1A mutation	6 (17%)	0 (0%)[e]	11 (17%)	23 (8%)
ARID1A mRNA (median)[f]	10.43	10.68[c]	10.56	10.4
ALB				
ALB mutation	3 (8%)	1 (3%)	6 (9%)	36 (13%)
ALB mRNA (median)[f]	15.65	19.34	19.58	20.44
AFP				
AFP mRNA (median)[f]	1.48	8.06[d]	12.97[e]	6.88[d]
HNF4A				
HNF4A mRNA (median)[f]	9.80[e]	10.84[e]	11.69	12.42
KRT19				
KRT19 mRNA (median)[f]	13.02[e]	11.78[e]	7.04	4.04

[a] Significant as compared to CAA.Like (p<0.05)

[b] Significant as compared to Blast.Like (p<0.05)

[c] Significant as compared to HCC (p<0.05)

[d] Significant as compared to CAA (p<0.05)

[e] Significant as compared to all other groups (p<0.05)

[f] mRNA expression is represented as Log2 median-centered values

Supplemental Table 2.8.3. Univariate and multivariate Cox regression analysis

Univariate analysis				
TCGA	Progression Free Interval	Progression Free Interval	Overall Survival	Overall Survival
Subtype (HCC ref)	Hazard Ratio (95% CI)	p – value	Hazard Ratio (95% CI)	p – value
CCA	1.22 (0.73 - 2.02)	0.448	1.90 (1.12 -3.23)	**0.017**
CCA-Like	1.68 (1.02 - 2.76)	**0.04**	1.45 (0.77 – 2.73)	0.245
Blast-Like	1.95 (1.32 – 2.88)	**<0.001**	3.72 (2.48 – 5.56)	**<0.001**
Stage (I ref)				
II	1.96 (1.38 – 2.77)	**<0.001**	1.63 (1.06 – 2.50)	**0.027**
III	2.35 (1.63 – 3.40)	**<0.001**	2.56 (1.69 – 3.87)	**<0.001**
IV	4.28 (2.20 – 8.35)	**<0.001**	5.26 (2.65 – 10.47)	**<0.001**
Grade (G1 ref)				
G2	1.44 (0.89 – 2.32)	0.133	1.34 (0.76 – 2.35)	0.316
G3	1.55 (0.95 – 2.52)	0.078	1.39 (0.78 – 2.48)	0.259
G4	1.51 (0.61 – 3.75)	0.775	2.46 (1.00 – 6.06)	**0.05**
Multivariate analysis				
Subtype + Stage (HCC ref)				
CCA	1.31 (0.79 – 2.18)	0.301	2.06 (1.21 – 3.51)	**<0.01**
CCA-Like	1.57 (0.95 – 2.59)	0.078	1.39 (0.73 – 2.63)	0.318
Blast-Like	1.70 (1.14 – 2.54)	**<0.01**	3.43 (2.28 – 5.17)	**<0.001**
Subtype + Stage +Grade (HCC ref)				
CCA	1.25 (0.72 – 2.11)	0.403	2.12 (1.22 – 3.68)	**<0.01**
CCA-Like	1.15 (0.85 – 2.46)	0.171	1.27 (0.64 – 2.50)	0.496
Blast-Like	1.62 (1.07 – 2.46)	**0.023**	3.30 (2.14 – 5.09)	**<0.001**
GSE14520				
Subtype (HCC ref)				
CCA-Like	1.49 (0.88 – 2.54)	0.142	1.60 (0.83 – 3.10)	0.161
Blast-Like	1.56 (1.06 – 2.30)	**0.025**	1.97 (1.24 – 3.14)	**0.004**
Stage (I ref)				
II	1.99 (1.30 – 3.05)	**<0.01**	2.07 (1.19 – 3.61)	**0.01**
III	3.17 (1.98 – 5.08)	**<0.001**	5.10 (2.92 – 8.93)	**<0.001**
Multivariate analysis				
Subtype + Stage (HCC ref)				
CCA-Like	1.1 (0.63 – 1.91)	0.745	1.11 (0.56 – 2.18)	0.77
Blast-Like	1.730 (0.87 – 1.95)	0.199	1.60 (.99 – 2.59)	0.054

Supplemental Table 2.8.4. Bonferroni corrected p-values of pairwise t-test between mutational signatures and subclasses

	S1	S2	S3	S4	S5	S6
CCA vs CCA-Like	1	1	1	1	0.26	0.95285
CCA vs Blast-Like	**0.002**	1	0.078	**0.00039**	**2.10E-11**	**0.00031**
CCA vs HCC	**0.0024**	1	0.093	0.16226	**2.10E-15**	6.40E-**09**
CCA-Like vs Blast-Like	0.0696	1	0.055	0.06847	**7.40E-05**	0.15424
CCA-Like vs HCC	0.1356	0.11	0.067	1	**2.40E-06**	0.**00075**
Blast-Like vs HCC	1	1	1	0.**00854**	1	0.39735

Bolded values denote p<0.05

CHAPTER 3: DNA DAMAGE REPAIR CLASSIFIER DEFINES DISTINCT GROUPS IN HEPATOCELLULAR CARCINOMA

3.1 Overview

Background: Hepatocellular carcinoma (HCC) prognosis remains dismal, with a 5-year survival rate of 18%. DNA repair pathways have been associated with variability in HCC clinical outcomes, but how DNA repair varies as a function of liver regeneration and other HCC characteristics is poorly understood.

Methods: We curated a panel of 199 genes representing fifteen DNA repair pathways to identify DNA repair expression classes and evaluate their associations with differentiation, mitotic and regenerative signatures and clinicopathologic variables in The Cancer Genome Atlas (TCGA) HCC study (n=374).

Results: We identified two groups in HCC, defined by Low or High expression across all repair pathways. The Low repair group had lower grade and retained expression of classical liver markers (*ALB, CYP450*), whereas the High repair group had more clinically aggressive features, increased p53 mutant-like gene expression, and high liver regenerative gene expression. These pronounced features overshadowed variation in the Low repair subset, but further heterogeneity was observed among Low repair samples. When considered separately, the Low repair group included three subgroups: L1, L2, L3. L3 had high DNA repair expression with worse progression-free (HR 1.24, 95% CI 0.81-1.91) and overall (HR 1.63, 95% CI 0.98-2.71) survival. High repair outcomes were also significantly worse compared to the L1 and L2 groups.

Conclusions: HCCs vary in expression of DNA repair pathways, and a subset of tumors with high regeneration have profoundly disrupted liver biology and poor prognosis.

3.2 Introduction

Hepatocellular carcinoma (HCC) is a heterogeneous cancer that varies vastly in clinical outcome and response to therapy. Although several studies have identified important molecular classes in HCC, there remains uncertainty regarding their associations with outcomes and their complex interaction with liver regenerative processes[11,87-90,93-95,119,130,175-177]. Recently, our lab defined three distinct subpopulations of HCC tumors, HCC, Blast-Like, and CCA-Like based on gene expression similarity to cholangiocarcinomas and hepatoblast cells[178]. The HCC class has prototypical HCC tumor characteristics. Blast-like has enrichment of *TP53* mutations and HBV positive status, exposure-related mutational signatures like the HCC class, and transcriptional patterns similar to hepatoblasts. While the CCA-Like resembles HCC histologically, but molecularly they looked more like cholangiocarcinomas with similar patterns of DNA mutations (*IDH1/2*, *BAP1*), DNA damage repair mutational signatures, and transcriptional patterns. These differences in key DNA-repair pathways suggest the importance of DNA repair, but exploration of DNA-repair specific signatures and their relationships to other cellular pathways could help elucidate their significance in HCC.

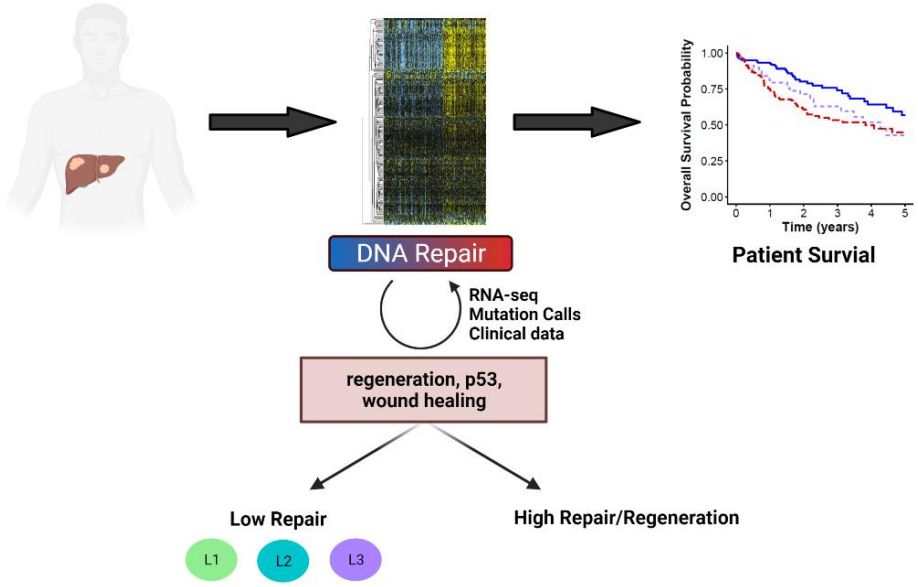
Hepatocytes, the chief functional cells of the liver, are responsible for many liver functions, such as detoxification, carbohydrate metabolism, lipid metabolism, and protein synthesis. If any of the regulatory pathways supporting liver function are impeded, the liver becomes more susceptible to advanced liver disease, including cirrhosis, hepatitis, and eventually, HCC[55]. These changes lead to increased rates of hepatocyte proliferation and impaired G_1/S checkpoint[179,180]. DNA repair plays a role in normal liver function and varies as a

function of the cell cycle[181]. Dysfunction in DNA repair pathways has been linked to many cancer types, but the liver's unique capacity to regenerate adds additional complexity to HCC.

It is well established that the liver's regenerative capability is tightly coupled to DNA repair processes[182-185]. Upon chronic liver damage, both processes are substantially dysregulated and there is an increased risk of genomic instability[109,180,186,187]. This dysregulation triggers a set of DNA damage response (DDR) pathways which orchestrate hepatocyte DNA repair, cell cycle arrest, and cell death. Aberrations of DNA repair and associated pathways, such as homologous recombination (HR), mismatch repair (MMR), and nonhomologous end joining (NHEJ), have been implicated in impairing liver genomic integrity, leading to activation of hepatocarcinogenesis and HCC development[188-191]. Given the importance of these two processes, they could have a joint impact on prognosis and therapeutic response. Specifically, dysregulation of DDR may determine chemoresistance, as shown in other tumor types[192,193].

Here, we investigated DNA repair defects in TCGA HCC study using a selected panel of 199 genes representative of fifteen DNA repair pathways (**Figure 3.2**). We sought to understand DNA repair in the context of liver homeostasis, including liver-specific gene expression, mitosis, and liver regeneration. We hypothesized that specific DNA repair patterns, along with liver regenerative capacity, could predict prognostic groups in HCC. Further characterizing the DNA repair pathways in both low and high repair HCC tumors, and integrating genomic information and clinical data, may help to better understand HCC heterogeneity in relation to outcomes.

Figure 3.2. Graphical abstract



3.3 Methods

3.3.1 Study population and datasets

Upper quartile normalized RSEM gene expression data for TCGA Hepatocellular carcinoma (HCC dataset, n=374) was downloaded from the GDC legacy archive (https://portal.gdc.cancer.gov /legacy-archive/) and log2 transformed. The data was median centered across samples.

Mutation data, including DNA variant allele frequency (VAF), was downloaded from Ellrott et al[194] (https://gdc.cancer.gov/about-data/publications/mc3-2017). Retrieved processed RNA data from TCGA PanCancer (https://gdc.cancer.gov/about-data/publications/pancanatlas)[195]. For each somatic mutation in the VAF file, we calculated RNA fragment allele counts. TCGA data is available through the GDC data portal, https://portal.gdc.cancer.gov/.

3.3.2 Classification of DNA repair groups in RNA expression data

We curated a list of 199 genes representing regulators of many DNA repair pathways (**Supplemental Table 3.8.1**). The panel was comprised of the following DNA repair pathways: Mismatch repair (MMR), Nucleotide Excision Repair (NER), Translesion Synthesis (TLS), Fanconi Anemia (FA), Base Excision Repair (BER), Nucleotide Metabolism, Template Switch, Poly ADP Ribose Polymerases (PARP), Checkpoint, DNA replication factors/Cell Cycle, Homologous Recombination (HR), Nonhomologous End Joining (NHEJ), Alternative End Joining (Alt-EJ), Cancer Testis Antigens (CTAs) [including HORMAD1 and MAGEA4, which are pathological cancer-specific activators of HR and TLS respectively[196,197]], and APOBEC cytosine deaminase family.

The 199-Repair score was calculated as the median of the DNA repair genes. We applied the R package mclust (v5.4.7)[198] to select the optimal cut-point, and based on the determined cut-point of 0.1, samples were classified as Low repair or High repair. Hepatitis B virus (HBV) was detected in HCC tumors RNAseq data via VirDetect[199]. Expression data was visualized using ComplexHeatmap (v2.8.0)[200] in R. All analyses was performed using R Statistical Software (v4.1.0, R Core Team 2021)[201] unless otherwise noted.

3.3.3 Define mitotic and regeneration patterns in HCC

We calculated the Mitotic index score using the median of the 10 mitotic pathway genes – BUB1, BUB1B, BUB3, CDC20, CDH1, ESPL1, MAD1L1, MAD2L1, PTTG1, TRIM69 – from the 199 DNA repair gene panel. A regeneration score from Colak et al[202] was derived from their regeneration activation and inhibition gene sets in by multiplying activation gene expression values by +1 and inhibition gene expression values by -1, then combined and the median value calculated. We compared continuous mitotic and regeneration score components by DNA repair groups using Wilcoxon signed-ranked tests.

3.3.4. Interrogation of biological states and processes in DNA repair groups

Gene Set Enrichment Analysis (GSEA)[155,156] was performed by comparing High repair group vs Low repair group for all HCC tumors, significance was determined by using a nominal p-value <0.05 and false-discovery rate (FDR) <0.25. We utilized the MSigDB Hallmark gene set and excluded any genes overlapping (103 genes) with our DNA repair gene panel to control for the fact the groups are derived from RNA expression data. Normalized Enrichment Scores (NES) were plotted for Hallmark pathways identified as significant for the High Repair and Low Repair group. We calculated pathway scores for selected significant pathways. Hallmark pathway scores and liver specific markers were compared stratified by Repair groups with adjacent normal using pairwise t-tests to investigate liver biological function.

3.3.5 Association of p53 and HRD

To assess the mutational landscape across the groups, we considered only genes mutated in at least 5% of on cohort to avoid bias due to genes mutated in a single sample and compared the differentially mutated genes between the groups using fisher test, then only significant genes (p<0.05) implicated in HCC were selected, *TP53*, *RB1*, *CTNNB1*. To interrogate defects in liver metabolic pathways, we specifically chose liver metabolic genes mutated in HCC >10%, yielding two genes: *APOB* and *ALB*. HRD mutations (*AXIN1, ATM, POLE, BRCA1, BRCA2, BARD1, BRIP1*) were chosen and examined based on the prognostic value in HCC identified by Chen et al[120]. We used a previously validated RNA signature that aggregates information on TP53-dependent genes to classify *TP53* functional status (mutant-like or wild-type) based on a similarity-to-centroid approach. *TP53* mutational status was classified based on presence (*TP53* mut) or absence (*TP53* WT) of somatic *TP53* mutations. We compared *TP53* RNA variant allele frequency (VAF) vs DNA VAF stratified by Repair groups using a linear model. Homologous recombination deficiency (HRD) scores, including

Loss of Heterozygosity (LOH)[203], large scale transitions (LST)[204], and number of sub chromosomal regions with allelic imbalance extending to the telomere (NtAI)[205], were extracted supplemental data from Knijnenburg et al[206]. Scores were dichotomized at a cut point of 20 based on the distribution of HRD scores in liver to ensure HRD status is tissue specific similar to prior literature[207-209].

We estimated the relative frequency differences (RFDs), representative of the difference between an index group and a reference group in the proportion of individuals exhibiting a given clinical or demographic feature, between the High Repair/Regenerative group (index group) and the Low repair group (referent group). RFDs and 95% confidence intervals (CI) were estimated using generalized linear models with binomials distributions and identity link functions.

3.3.6 Low Repair tumor classification

To unmask differences in the Low repair group, we removed samples classified in the High Repair/Regenerative group from the TCGA HCC expression data. We performed clustering analysis using the R package ConsensusClusterPlus (v1.56.0)[134] on the Low repair group to evaluate expression differences present within the Low repair group.

3.3.7 Clinical variables and survival analysis

We extracted TCGA clinicopathologic and survival data from Liu et. al[148]. Clinical data was filtered to samples with ≥94% of available clinical annotations across the study for race, gender, age, pathologic stage, pathologic primary tumor (pT), grade, and survival. AJCC TNM Classification 2010 (7th Edition)[210] was used for pathologic primary tumor (pT) and samples classified by AJCC TNM Classification 6th Edition[211] were converted to the 7th edition. Samples classified based on AJCC 4th and 5th edition TNM staging system[212,213] were excluded from analyses due to lack of sufficient information for conversion to the 7th edition. Kaplan-Meier curves with log rank tests were generated using the R package survminer (v0.4.9)[147]. Vascular invasion

was grouped by presence of Univariate Cox regression was used to determine significance of DNA repair groups, sex, race, molecular subtype, stage, grade, and HBV status. Models were adjusted for pathological tumor stage and tumor grade to determine whether associations between repair status and survival held. Continuous variable comparisons were made by Repair groups using Wilcoxon signed-rank test or ANOVA as indicated. Statistical analyses were performed using R unless otherwise noted.

3.4 Results

3.4.1 HCC tumors exhibit two groups based on expression of 199 DNA repair genes

To characterize DNA repair patterns in TCGA hepatocellular carcinoma (HCC) samples, we calculated an RNA-based DNA repair score based on 199 DNA repair genes and identified an optimal cut point to distinguish the groups. HCC tumors were classified into two groups based on DNA repair gene expression: Low repair (n=216) and High repair (n=158) (**Figure 3.7.1a**). The High repair group includes many upregulated genes across all DNA repair pathway, pointing to high activity of DNA repair genes. High repair significantly associated with Blast-like and CCA-like molecular subtypes, Asian race, enrichment of HBV positive cases, and higher tumor stage, pathologic primary tumor (pT), and grade (**Table 3.7.1**). The median age at diagnosis was significantly lower in the High repair group. There were no significant differences in vascular invasion between the repair groups. Low repair samples represent prototypical HCC tumors as majority were classified in the HCC molecular subtype and were significantly enriched for lower stage and grade tumors.

Based on overall high DNA repair expression seen in the High repair group, we hypothesized that this group may have high activity of mitotic and liver regenerative pathways. To investigate, we examined mitotic and liver regenerative pathway patterns stratified by the Repair groups. Mitotic index score and regenerative score were significantly increased in the

High repair group compared to the Low repair group (p<0.001, **Figure 3.7.1b-c**). Liver regeneration-associated repair patterns neglect biological heterogeneity.

To further interrogate biological processes occurring in these groups, we performed gene set enrichment analysis (GSEA) using the Hallmark gene signature set to identify differential pathway signatures between the High and Low repair group (**Supplemental Figure 3.8.1a**). The High repair group was significantly enriched for 5 pathways including cell cycle gene sets (G2M checkpoint, E2F targets, Mitotic spindle), spermatogenesis and MYC targets. The Low repair group had 8 significantly enriched pathways all related to liver biology and function (Adipogenesis, xenobiotic metabolism, fatty acid metabolism, coagulation, bile acid metabolism, oxidative phosphorylation, peroxisome, and reactive oxygen species). The adipogenesis and fatty acid metabolism signature scores were significantly higher in the Low repair but still not as high as the tumor adjacent normal tissue (p<0.001 for all pairwise comparisons, **Supplemental Figure 3.8.1b-c**). MYC target genes were significantly upregulated in the High repair samples compared to the Low repair (p<0.001, **Supplemental Figure 3.8.1d**). The High repair group displayed significantly higher expression of cell cycle and hepatoblast marker *AFP* compared to the Low Repair group and Tumor adjacent normal (p<0.001 for all pairwise comparisons, **Supplemental Figure 3.8.1e**). The Low repair group also had significantly increased expression of liver markers *ALB* and cytochrome P450 (*CYP450*), but the expression levels were still not as high as normal liver (p <0.01, **Supplemental Figure 3.8.1f-g**). The Low repair group displays a more liver biology gene expression, while the High repair group shows decreased liver gene expression and increased DNA repair dysregulation.

3.4.2 High Repair classes associate with p53 functional status and TP53 mutation status

Given ~36% of HCC tumors have *TP53* mutations and the mutual regulation between cell cycle and p53, we hypothesized p53 plays a vital role in liver repair dysfunction and

genomic instability. We assessed the mutational landscape based on the Repair groups. Of 781 recurrently mutated genes present in at least 5% of liver samples, only *TP53* and *CTNNB1* were significantly different between the repair groups. The High repair group had a significantly higher rate of *TP53* mutations (44.9%) compared to the Low repair group (19%, p<0.001, **Table 3.7.2**). Interestingly, *CTNNB1* mutations were significantly enriched in the Low repair group (32.4%) compared to the High repair group (18.4%) (p<0.001). *CTNNB1* mutations occur frequently in HCCs, leading to activation WNT/β-catenin signaling in 30-50% of HCC cases[12]. Another key liver gene *ALB* was more frequently mutated in the Low repair group (19.4%) than the High repair group (11.4%) though did not reach significance. The key HRD genes (*ATM*, *POLE*, *BRCA1/2*, *BARD1*, *BRIP1*) were less than 5% mutation rate in both groups, except for DNA damage response pathway gene *AXIN1* which was higher in the High repair (9.5%) as compared to the Low Repair (6.0%). Based on the presence HRD mutated genes by repair groups, HRD genes were mutated at a higher frequency in the High repair (21.5%) than the Low repair (17.6%).

To further consider the role of master regulator *TP53*, we used a previously validated RNA signature to classify tumor for p53 functional status (mutant-like/Wild-Type [WT]) [214]. Patterns of expression for the p53 gene signature are shown across the HCC tumors in **Figure 3.7.2a**. Two groups are evident: one enriched for p53 mutant-like and the other enriched for p53-Wild-Type (WT). The High Repair group was significantly enriched for tumors classified as p53 mutant-like, accounting for 75% of the p53 mutant-like HCC samples (p<0.001, **Supplemental Table 3.8.2**).

We evaluated whether p53 status (RNA), *TP53* mutational status (DNA), HRD and HBV status were associated with the DNA Repair groups. The High Repair group had enrichment for p53 mutant-like status (RNA), *TP53* mutation status (DNA), HRD High status and HBV positive status (**Figure 3.7.2b, Supplemental Table 3.8.2**). In the High repair group, the mutant p53

signature (RFD: 66.5%, 95% CI 58.5-73.5) was even more prevalent than *TP53* DNA mutation status (RFD: 26.0%, 95% CI 16.6-35.2) (**Supplemental Figure 3.8.2a**). Increased HBV positive status tracks with the correlation of increased TP53 mutations in HBV+ HCC tumors. High repair tumors had significantly higher HRD scores than Low repair tumors ($p<0.001$). HRD high tumors were unequally distributed between Low repair and High repair tumors (RFD: 34%, 95% CI 24.2-44.0) (**Supplemental Figure 3.8.2b**). High repair samples are more likely to be p53 mutant-like as well as high DNA variant allele frequency (VAF) suggesting mutation and single copy loss or loss of heterozygosity and high RNA expression of the mutant allele (**Supplemental Figure 3.8.2c**). These results imply there is strong *TP53* dysfunction in the High repair group. Differences in p53 mutant status underscores the importance of p53 in liver tissue homeostasis.

3.4.3 Clustering analysis reveals three subgroups within Low repair groups

The overall, general high expression of DNA repair genes in the High repair/Regenerative group may mask heterogeneity within the Low Repair group. To assess this, we performed consensus clustering in only the samples classified as Low repair. The clustering analysis revealed three subgroups – L1, L2, L3, with increased heterogeneity across the DNA repair genes (**Figure 3.7.3a**). L1 was significantly decreased for DNA repair activity compared to L2 and L3, but L3 showed greater similarities to the higher repair activity seen in the High repair group (**Figure 3.7.3b**, L1 vs rest $p<0.001$). L1 and L3 were defined by higher expression of homologous recombination (HR) related genes (*BRIP1, BRCA2, BARD1, RAD51AP1*). L2 and L3 were characterized by higher expression of cell cycle and mitotic checkpoint genes (*BUB1* and *BUB1B*). L2 showed higher expression of replication factors (*POLD1, RB1,* and *HORMAD1)*. When we examined associations with clinicopathological feature and risks between the subgroups, only race, grade, and HBV status were significant (**Table 3.7.3**). L1 was associated with lower grade while L2 and L3 were significantly enriched

for higher grade tumors (p<0.001). Asian ancestry and HBV positive status was increased in L2 and L3 compared to L1 (p=0.002 [p-value based on Asian vs White], p=0.001). There were no associations with pathologic tumor stage, pathologic primary tumor stage (pT), or vascular invasion across the Low repair groups. We assessed differences in *TP53* functional status, *TP53* mutational status, and HRD status in the Low repair subgroups (**Supplemental Table 3.8.3**). L2 and L3 were significantly enriched for p53 mutant-like (57.8% and 42.2%, respectively) and *TP53* mutations (29.3% and 61.0%, respectively) which were mostly lacking in L1 (p53 mutant-like 0%, TP53 mutation 9.7%, p<0.001). Also, L2 and L3 had significantly higher HRD scores than L1 (p<0.001).

3.4.4 High Repair group has worse overall survival and progression-free survival

We investigated progression-free and overall survival differences between the repair groups. Clinical outcomes were similar for L1 and L2 and were combined for analyses. (**Figure 3.7.4**). High repair samples have significantly worse progression-free (Hazard Ratio [HR] 1.78, 95% CI 1.26-2.49) and overall (HR 1.97, 95% CI 1.29-2.99) survival compared to L1+L2 tumors. Similarly, the L3 group has significantly worse than progression-free (HR 1.24, 95% CI 0.81-1.91) and overall (HR 1.63, 95% CI 0.98-2.71) survival outcomes. Both L3 and High repair tumors retained significance for worse overall survival, but only the High repair tumors retained significance for worse progression free survival when pathologic stage and grade were included in the model.

3.5 Discussion

We used a curated gene panel to identify DNA repair expression classes in TCGA HCC tumors. We identified two DNA repair groups (Low repair and High repair/Regenerative) with distinct biological patterns. The Low repair group was characterized by classical HCC tumor features and lower grade, while the High repair group had high expression of all DNA repair

genes, associated with Blast-like and CCA-like molecular subtypes, higher stage, Asian race, and worse prognosis. The High repair group overshadowed subtleties in the Low repair, and therefore, clustering solely in the Low repair group elucidated three subgroups: L1, L2, L3. The L1 and L2 subgroups were characterized by liver pathways and lacked regenerative processes associated with the High repair group. L3 retained expression of liver-related genes, but was more similar to DNA repair expression and survival outcomes of the High repair group potentially representing a transitory state that could shift to a High repair phenotype.

Previous studies have detected similar molecular classes, most studies emphasized two main groups based on DNA repair RNA expression[93-96]. Three previous studies identified two DNA repair groups: Oshi et al. found a low and high repair group based on a gene set that included 150 genes[94], P. Lin et al. classified a DNA repair activated and suppressed group using 276 genes[95], and Chen et al. found a low and high risk DNA repair group based on 23 DDR-related gene pairs[96]. Across these studies, similar to our results the high repair activated samples were associated with worse survival, and these authors further found high repair/DDR-activated samples were associated with distinct immune profiles, poor differentiation, elevated intratumor heterogeneity and mutation burden. Of these three studies, gene overlap with our 199-gene set varied (Overlap was 26, 122, and 11 genes for Oshi, P. Lin, and Chen, respectively). The similarities observed across these studies, despite the different number of overlapping genes, underscores that these high repair patterns are robust and can be detected with few features. One advantage of our gene list, which covers many DNA repair pathways, was that we included several essential DNA repair genes such as *BUB1, BUB1B, RAD51AP1, RB1*, and *HORMAD1* that are known to have significant clinical relevance[215-220] and helped us to define the heterogeneity among the low DNA repair group. Mitotic and replicative factors, along with HR expression, helped to distinguish low repair groups. While there is agreement among groups on the two-classification schema, extending beyond two groups has proven more difficult

with little resolution. For example, one lab found four groups using a combination of manual and cluster-based methods. Whereas we used a data-driven clustering approach, and while both studies found four groups, there were different gene sets utilized and the groups varied greatly. It is clear regardless of the size of the gene set, we can identify the High repair groups, but more resolution is needed to identify the low repair samples.

Liver regeneration is a key liver feature that plays a role in HCC development by supporting a tumorigenic environment[221,222]. We found elevated expression of liver regeneration signatures in the High repair group. High repair/regenerative tumor phenotypes were related to high grade and Blast-like molecular subtype. Samples classified within the High repair group showed increased wound healing (i.e., increased mitotic activity), increased regenerative activity, increased p53 dysfunction, indicating these processes act in tandem. Many of these pathways have been implicated in micrometastases and *de novo* cancers following surgical resection for liver diseases and HCC[221,223-225], emphasizing that these pathways are interconnected and necessary for liver homeostasis.

We further contextualized DNA repair in association with other established pathways in implicated in HCC (p53, HBV status, liver inflammation) to reveal underlying dysregulation. Prevalent p53 dysfunction is a hallmark of the High repair group. While *TP53* mutations were enriched in this group, an RNA signature of pathway dysfunction was observed in almost every single High repair sample. Due to p53s role in mediating senescence post liver injury, it is responsible for regulation of fibrosis and may prevent deterioration to HCC[226]. In addition, mitotic activity is tightly and negatively controlled by p53[227], mitotic dysregulation is strongly associated with improper cell division and aneuploidy[64], and may promote inflammatory environments[228-230].

With respect to HBV, the TCGA includes representation from Southeast Asian patients where a main risk factor for HCC is HBV infection. HBV subverts DNA damage repair (DDR)

pathways creating an inflammatory environment that promotes hepatocarcinogenesis and HCC development. Data has linked chronic HBV infection to immune induced liver injury[168-170], this injury can result in hyperactive DDR, and in turn lead to a more aggressive chemo-resistant unresectable liver cancer. We also found that HBV positive HCC tumors associated with increased expression of cell cycle and mitotic genes. While viral hepatitis vs nonviral classification paradigm is an important clinical distinction and we do observe enrichment of HBV infection status in the High Repair group, there still was heterogeneity in the DNA repair groups across HCC and outcomes may not be solely driven by HBV infection status. We also had limited information for risk factors such as nonalcoholic fatty liver disease (NAFLD) and non-alcoholic steatohepatitis (NASH). This emphasizes the importance of integration of risk factors and molecular features versus stratification by risk factors.

The major advance of our work was in showing that separating high repair/regenerative tumors allowed improved characterization of heterogeneity among Low repair HCCs. This helped identify the L3 tumor group, which has some preserved liver biology-related gene expression, but higher HRD and DNA repair dysfunction. L3 also had higher frequency of samples with a p53 mutant-like call and *TP53* mutations. While the L2 group was enriched with tumors that were higher grade and patients with Asian ancestry and HBV infection compared to L1, there were no differences in survival outcomes between the groups. By stratifying on repair status, we identified classes that predict outcomes. Regardless of grade and stage, L3 and High repair had worse prognosis compared to the L1 and L2 tumors. TCGA has a short length of follow-up and lack of liver disease scoring (Child-Pugh classification, Ishak fibrosis stage) and treatment data, limiting our ability to investigate the role of our DNA repair classifiers in evaluating extent of liver damage and HCC patient therapy. More comprehensive and standardized annotation for clinical data elements will be essential to better understand associations between DNA repair defects, liver regenerative mechanisms, and risk factors in HCC.

In summary, this work expands on prior studies to identify four DNA repair classes in HCC (L1, L2, L3, and High repair/Regenerative) that are associated distinct biology and clinical prognosis. Our data suggests that DNA repair and liver regeneration work in tandem to define HCC heterogeneity. Future work should evaluate heterogeneity in DNA repair in association with specific chemotherapeutic regimens to address the need for improved treatment strategies for this poor prognosis cancer type.

3.6 Acknowledgements

We thank the members of the K.A.H. and M.A.T laboratories for useful discussions. **Figure 3.2** was created in biorender.com. This work was supported by National Institutes of Health Grant 5 U24 CA210988 (to K.A.H) and UNC-CH Cancer Control Education Program (T32CA057726, to A.W.).

3.7 Figures and Tables

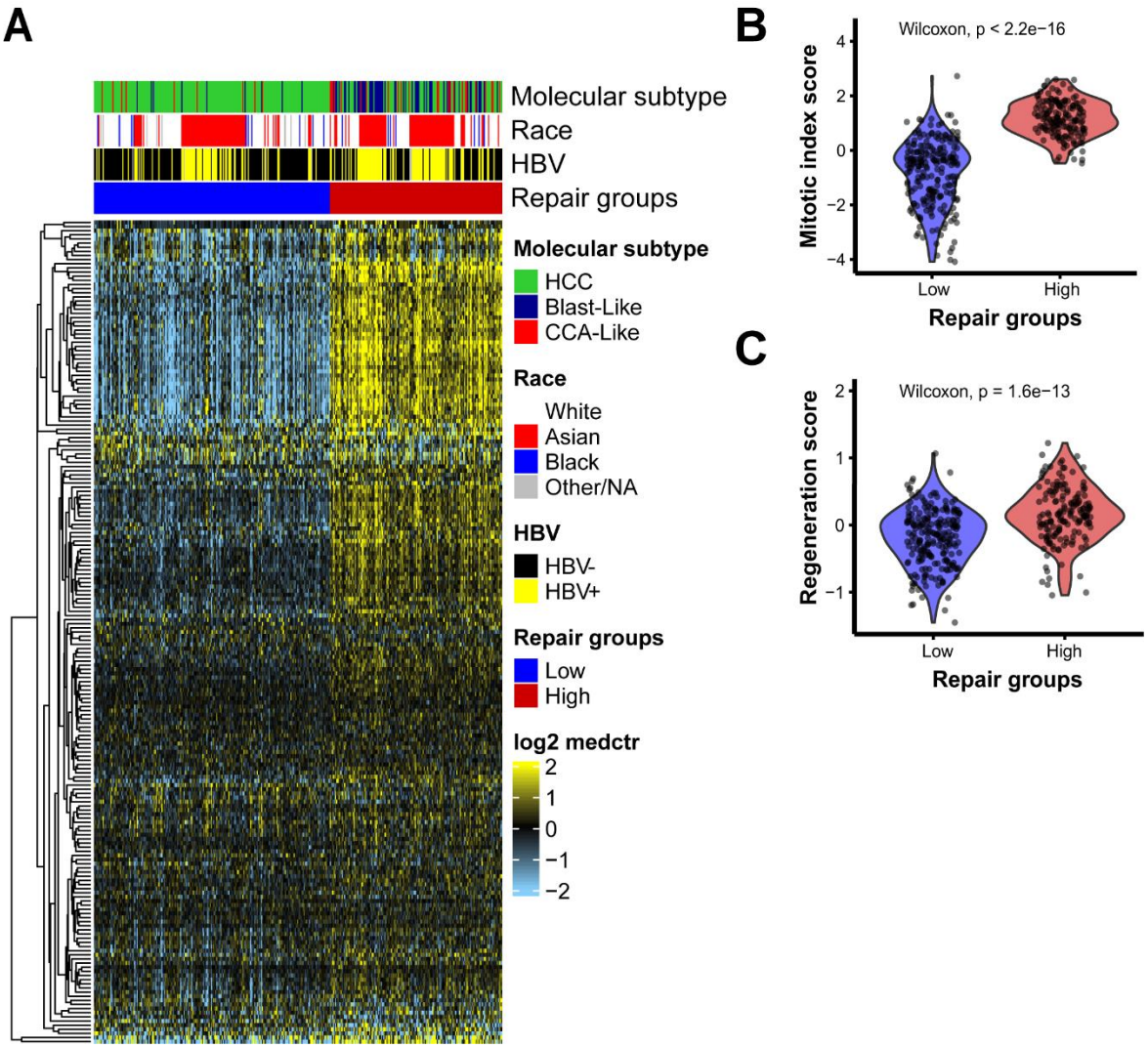


Figure 3.7.1. HCC tumors separate into two distinct groups based on DNA repair pathway gene expression

(a) Expression heatmap of 199 DNA repair genes ordered by Repair groups. The heatmap scale is low expression (light blue) to high expression (yellow). Rows represent DNA repair genes and columns samples. Annotation tracks show molecular subtype, race, HBV status, and Repair groups. Violin plot of (b) mitotic index score and (c) regeneration score by Repair groups, Low repair (blue) and High repair (red), p-value for Wilcoxon test between groups.

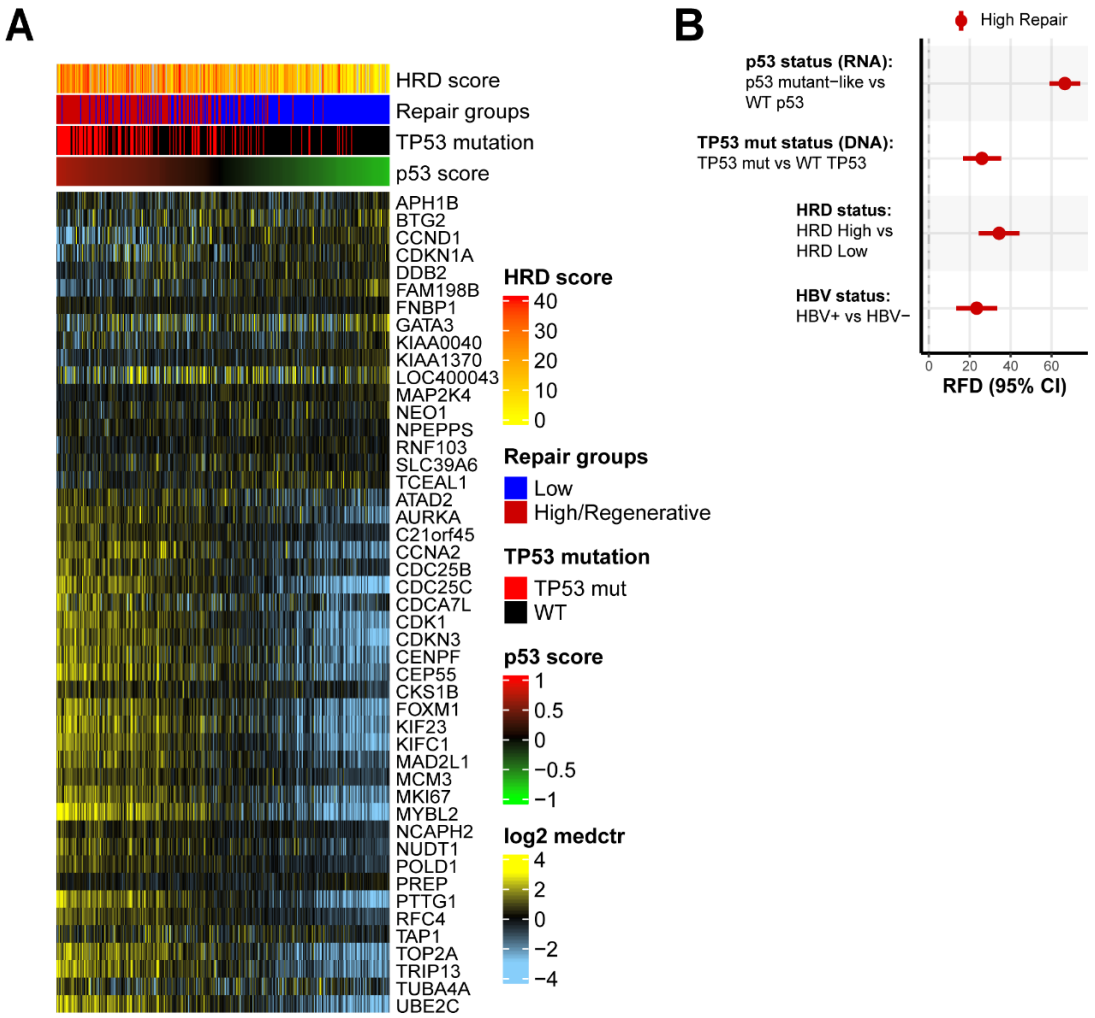


Figure 3.7.2. High Repair classes associate with dysfunctional p53 functional status and TP53 mutation status

(a) Expression patterns of p53 gene signature with samples ordered by p53 score and sorted by p53 mutant-like genes. The heatmap scale is low expression (green) to high expression (red). Rows are DNA repair genes and columns are samples. Annotation tracks show HRD score, Repair groups, *TP53* mutation status, and p53 score. (b) Relative Frequency Difference (RFD) analysis of features in High group compared to Low Repair group. Features include p53 mutant status, TP53 mutation status, HRD status, and HBV status. 95% confidence intervals (CI) are included for each measure.

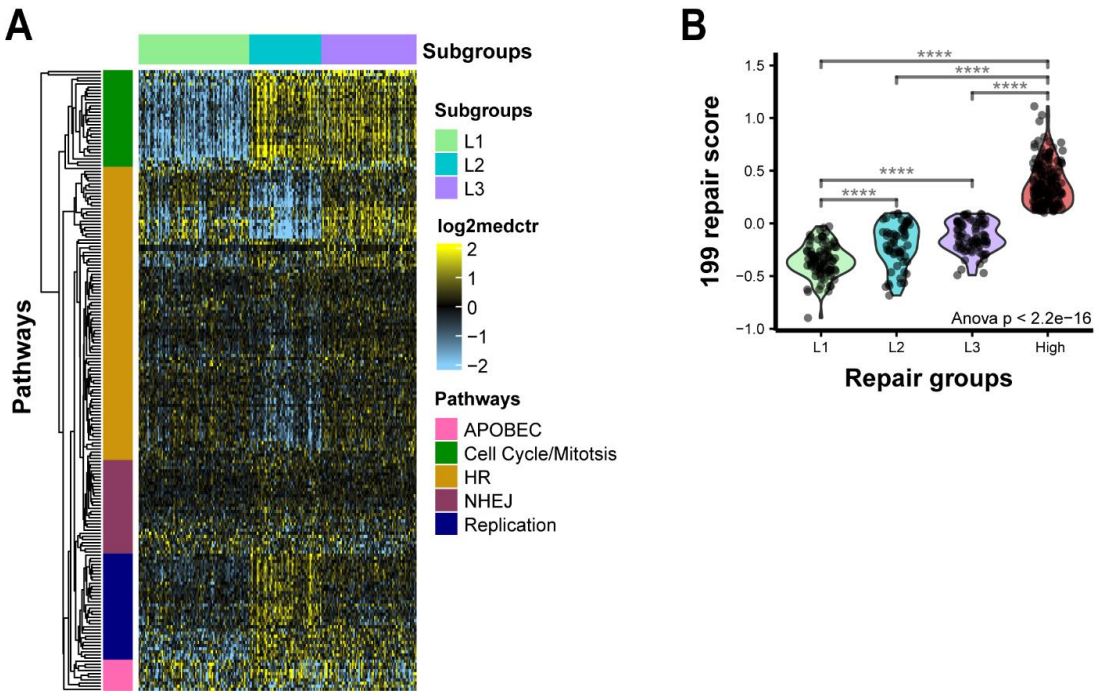


Figure 3.7.3. DNA repair pathway gene expression heterogeneity within the Low repair group

(a) Heatmap DNA repair signatures in Low repair group. The heatmap scale is low expression (light blue) to high expression (yellow). Rows are DNA repair genes and columns are samples. Annotation track shows clusters: L1 (light green), L2 (turquoise), L3 (purple). Major gene pathways in each gene cluster are annotated on the rows: APOBEC (pink), Cell cycle/mitosis (green), HR (gold), NHEJ (plum), Replication (navy).

(b) Repair gene score identifies distinct features across early and advanced HCC tumors. Violin plot of repair score by repair groups, **** indicates p<0.0001 for two sample t-test between High expression group and L1.

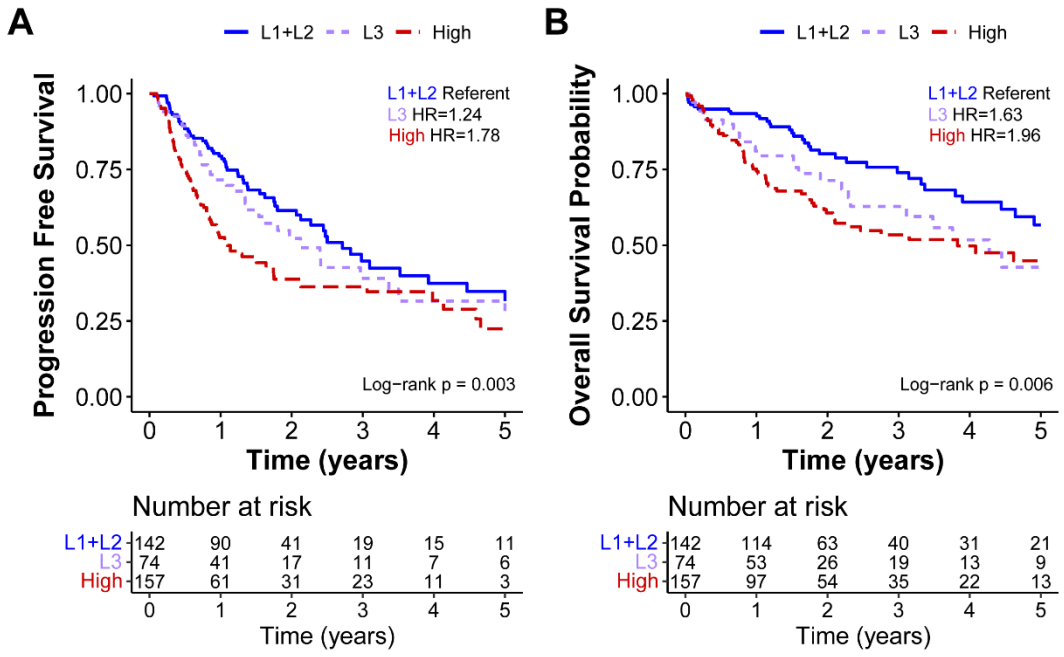


Figure 3.7.4. Low repair subgroups L1 and L2 have better progression-free and overall survival

Kaplan-Meier curves of TCGA HCC data for (a) progression-free and (b) overall survival. All survival data was censored at 5 year, and hazard ratios, 95% CI, and log-rank p-values were calculated for each measure.

Table 3.7.1. Overall patient characteristics of TCGA HCC study, overall and stratified by 199-Repair groups.

	Overall	Low	High	p-value	RFD [95% CI]
n	**374**	**216**	**158**		
Age at diagnosis					
Mean ± SD	59.48 (13.47)	60.74 (± 13.89)	57.75 (± 12.71)	0.034	-
Sex					
Female	121 (32.4)	65 (30.1)	56 (35.4)	0.327	REF
Male	253 (67.6)	151 (69.9)	102 (64.6)		-5.35 [-15.0 – 4.3]
Race[a]					
White	185 (51.1)	117 (57.1)	68 (43.3)	0.025	REF
Asian	160 (44.1)	78 (38.0)	82 (52.2)		14.2 [3.92 – 24.4]
Black	17 (4.7)	10 (4.9)	7 (4.5)		-0.42 [-4.79 – 3.95]
Molecular Subtype					
HCC	275 (73.5)	198 (91.7)	77 (48.7)	<0.001	REF
Blast-Like	66 (17.6)	9 (4.2)	57 (36.1)		31.9 [24.0 – 39.9]
CCA-Like	33 (8.8)	9 (4.2)	24 (15.2)		11.0 [4.82 – 17.2]
AJCC Pathologic Tumor Stage[b]					
I	173 (49.4)	116 (57.7)	57 (38.3)	0.001	REF
II	87 (24.9)	44 (21.9)	43 (28.9)		6.97 [-2.28 – 16.2]
III/IV	90 (25.7)	41 (20.4)	49 (32.9)		12.5 [3.11 – 21.9]
AJCC Pathologic Primary Tumor (pT)[c]					
T1	178 (51.4)	120 (61.9)	58 (38.2)	<0.001	REF
T2	88 (25.4)	40 (20.6)	48 (31.6)		11.0 [1.68 – 20.3]
T3/T4	80 (23.1)	34 (17.5)	46 (30.3)		12.7 [3.74 – 21.9]
Grade[d]					
G1/G2	233 (62.3)	155 (73.1)	78 (49.7)	<0.001	REF
G3/G4	136 (36.4)	57 (26.9)	79 (50.3)		23.4 [13.6 – 33.3]
Cirrhosis/Fibrosis[e]					
None	75 (34.9)	49 (36.6)	26 (32.1)	0.604	REF
Cirrhosis/Fibrosis	140 (65.1)	85 (63.4)	55 (67.9)		4.47 [-8.56 – 17.5]
HBV infection[f]					
Negative	217 (59.0)	146 (68.9)	71 (45.5)	<0.001	REF
Positive	151 (41.0)	66 (31.1)	85 (54.5)		23.4 [13.4 – 33.4]
Vascular invasion[g]					
None	208 (65.4)	115 (65.0)	93 (66.0)	0.948	REF
Micro/Macro	110 (34.6)	62 (35.0)	48 (34.0)		0.98 [-9.57 – 11.4]

[a] Excludes samples noted as other and NA (n=12)

[b] Excludes samples without pathologic stage annotation (n=24)

[c] Excludes samples based on AJCC 4th and 5th edition and noted as NA (n=28)

[d] Excludes samples without grade annotation (n=5)

[e] Excludes samples without annotation (n=159)

[f] HBV infection determined by >5 HBV reads from RNA-seq

[g] Excludes samples without annotation (n=56)

Table 3.7.2. Overall mutation rate for frequently mutated, liver metabolic and DNA repair genes in HCC stratified by 199 Repair groups

Gene	Low Repair (n=216)	High Repair (n=158)	
Frequently mutated genes			
TP53	41 (19.0%)	71 (44.9%)	***
CTNNB1	70 (32.4%)	29 (18.4%)	***
Liver metabolic mutated genes			
ALB	42 (19.4%)	18 (11.4%)	ns
HRD mutated genes			
AXIN1	13 (6.00%)	15 (9.50%)	ns
ATM	7 (3.24%)	7 (4.43%)	ns
POLE	4 (1.85%)	1 (0.63%)	ns
BRCA1	4 (1.85%)	1 (0.63%)	ns
BRCA2	6 (2.80%)	2 (1.30%)	ns
BARD1	2 (0.93%)	5 (3.20%)	ns
BRIP1	3 (1.40%)	3 (1.90%)	Ns

*** indicate p<0.001

Table 3.7.3. Clinicopathological features and risk factors of 3 subgroups in the Low Repair group

	L1	L2	L3	p-value
n	**86**	**56**	**74**	
Age				
Mean (± SD)	60.30 (± 15.04)	58.23 (± 13.84)	63.15 (± 12.23)	0.126
Gender				
Female	24 (36.9)	18 (27.7)	23 (35.4)	0.843
Male	62 (41.0)	38 (25.2)	51 (33.8)	
Race[a]				
White	56 (47.9)	21 (17.9)	40 (34.2)	0.002*
Asian	21 (26.9)	29 (37.2)	28 (35.9)	
Black	5 (50.0)	0 (0.0)	5 (50.0)	
Molecular subtype				
HCC	81 (40.9)	50 (25.3)	67 (33.8)	0.192
Blast-Like	1 (11.1)	5 (55.6)	3 (33.3)	
CCA-Like	4 (44.4)	1 (11.2)	4 (44.4)	
AJCC Pathologic Tumor Stage[b]				
I	45 (38.8)	27 (23.3)	44 (37.9)	0.446
II	17 (38.6)	16 (36.4)	11 (25.0)	
III/IV	16 (39.0)	12 (29.3)	13 (31.7)	
AJCC Pathologic Primary Tumor (pT)[c]				
T1	48 (64.0)	26 (54.2)	46 (64.8)	0.691
T2	15 (20.0)	13 (27.1)	12 (16.9)	
T3/T4	12 (16.0)	9 (18.8)	13 (18.3)	
Grade[d]				
G1/G2	74 (47.7)	33 (21.3)	48 (31.0)	<0.001
G3/G4	10 (17.5)	23 (40.4)	24 (42.1)	
Cirrhosis/Fibrosis[e]				
Cirrhosis/Fibrosis	31 (36.5)	20 (23.5)	34 (40.0)	0.028
No	27 (55.1)	13 (26.5)	9 (18.4)	
HBV status[f]				
Negative	69 (47.3)	30 (20.5)	47 (32.2)	0.001
Positive	15 (22.7)	26 (39.4)	25 (37.9)	
Vascular invasion[g]				
None	21 (29.6)	18 (39.1)	23 (38.3)	0.46
Micro/Macro	50 (70.4)	28 (60.9)	37 (61.7)	

[a] Excludes samples noted as other and NA (n=9)

[b] Excludes samples without pathologic stage annotation (n=15)

[c] Excludes samples based on AJCC 4th and 5th edition and noted as NA (n=22)

[d] Excludes samples without grade annotation (n=4)

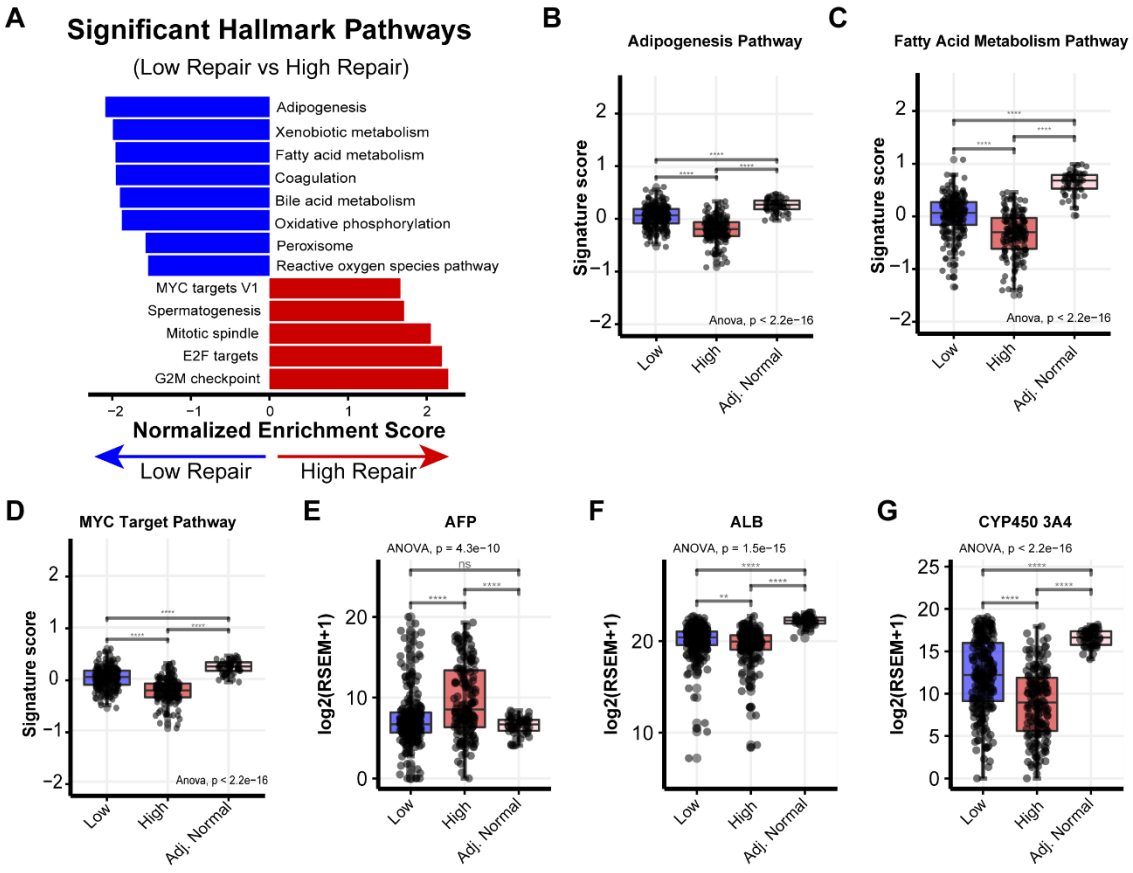
[e] Excludes samples without annotation (n=82)

[f] HBV infection determined by >5 HBV reads from RNA-seq

[g] Excludes samples without annotation (n=39)

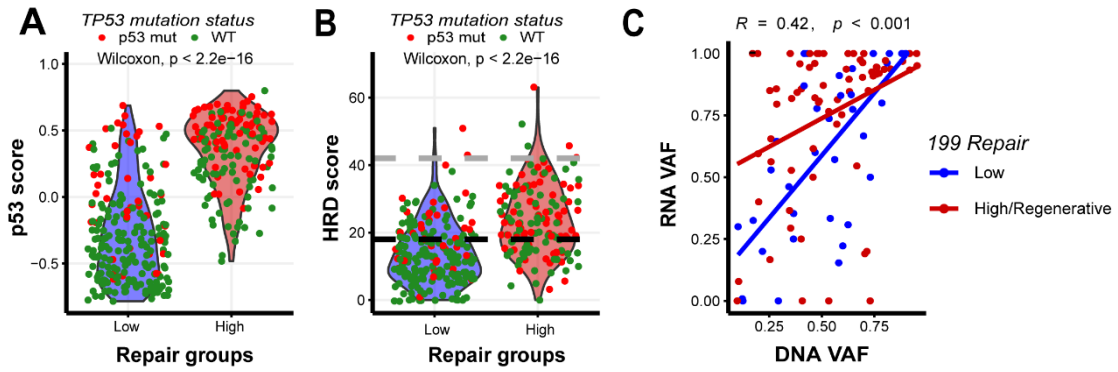
*p-value based on Asian vs white people

3.8 Supplemental Figures and Tables



Supplemental Figure 3.8.1. Low repair samples display expression of genes involved in liver function

(a) Gene Set Enrichment Analysis was performed Low Repair (blue) vs High Repair (red) using the Hallmark gene sets. Significant pathways were plotted by normalized enrichment score (NES). One-way ANOVA p-value is displayed. (b) Adipogenesis, (c) Fatty acid metabolism, and (d) MYC target pathway scores are plotted by Repair groups and Tumor adjacent normal. (e-g) Gene expression for hepatoblast marker *AFP*, hepatocyte markers *ALB*, and liver metabolomic marker *CYP450* are shown by Repair group. For pairwise comparisons between groups, ns indicates not significant, *** indicates p<0.001 and **** indicates p<0.0001 for all pairwise t-test.



Supplemental Figure 3.8.2. TP53 dysfunction and high HRD are defining features of High repair tumors

(a) Violin plot of p53 score by Repair groups. Points are colored by TP53 mutation status, Low repair (blue) and High repair (red); p < 2.2e-16 for Wilcoxon test between repair groups.

(b) Violin plot of HRD score by Repair groups. Points are colored by TP53 mutation status, Low (blue) and High (red); p < 2.2e-16 for Wilcoxon test between repair groups.

(c) Variant allele frequencies for *TP53* were calculated from the RNA and DNA sequencing reads. Repair groups are indicated by point color (blue=Low and red=High/Regenerative). Regression lines are plotted and colored according to Repair group.

Supplemental Table 3.8.1. DNA repair genes by pathway

Pathway	Error Propensity	Genes
Mismatch Repair	Low	EXO1, MLH1, MLH3, MSH2, MSH3, MSH6, PMS2
Nucleotide Excision Repair	Low	ERCC1, ERCC2, ERCC3, ERCC4, ERCC5, ERCC6, LIG1, LIG3, RAD23B, XPA, XPC, XRCC1
Trans-Lesion Synthesis	Low	MAD2L2, POLH, POLI, POLK, RAD18, REV1, UBE2A, UBE2B, USP7
Fanconi Anemia	Low	BTBD12, C17orf70, C1orf86, FANCA, FANCB, FANCC, FANCD2, FANCE, FANCG, FANCI, FANCL, FANCM, MTMR15, PALB2, UBE2T, USP1
Base Excision Repair	Low	APEX1, FEN1, NEIL1, NEIL2, NEIL3, OGG1, PCNA, PNKP
Nucleotide metabolism	Low	RNASEH1, RNASEH2B, RRM1, RRM2, RRM2B, TK1
Template Switch	Low	SHPRH, ZRANB3
Poly ADP Ribose Polymerases	Low	C12orf48, PARP1, PARP2, PARP3, PARP4, PARP9
Checkpoint	High	ATM, ATR, CHEK1, CHEK2, CLSPN, HUS1, HUS1B, MDC1, RAD1, RAD9A, RAD9B, RAD17, TIMELESS, TIPIN
DNA replication factors/Cell Cycle	High	AURKA, AURKB, CCNA1, CCNA2, CCNB1, CCND1, CCND2, CCND3, CCNE1, CDC25A, CDC25B, CDC25C, CDC45, CDC6, CDKN1A, CDKN1B, CDKN2A, CDT1, DDB1, DDB2, E2F1, E2F2, E2F3, E2F4, E2F5, E2F6, E2F7, GMNN, MAX, MCM10, MCM7, MYC, PHB, PLK1, POLD3, POLD4, POLE4, PPP1R12A, RB1, RBL1, RBL2, RFC2, RFC3, RFC4, RFC5, RPA1, RPA2, RPA3, WEE1
Homologous Recombination	High	BARD1, BLM, BRCA1, BRCA2, BRIP1, DNA2, FAM175A, FBXO5, MUS81, RAD51, RAD51AP1, RAD51C, RAD51L1, RAD51L3, RAD54B, RAD54L, RAD54L2, RBBP8, RNF138, TOP3A, TOP3B, TOPBP1, UIMC1, XRCC2
Nonhomologous End Joining	High	DCLRE1C, DNTT, LIG4, MRE11A, NBN, NHEJ1, POLB, POLL, POLM, POLQ, RAD50, RIF1, RNF168, RNF169, RNF8, TP53BP1, XRCC4, XRCC5, XRCC6
Mitosis/Spindle Assembly Checkpoint	High	BUB1, BUB1B, BUB3, CDC20, CDH1, ESPL1, MAD1L1, MAD2L1, PTTG1, TRIM69
Repair/Replication DNA Polymerases	High	POLD1, POLE
APOBEC		APOBEC1, APOBEC3A, APOBEC3B, APOBEC3C, APOBEC3D, APOBEC3F, APOBEC3G, APOBEC3H
Cancer Testes Antigens		HORMAD1, MAGEA1, MAGEA4, MAGEA6, MAGEA10

Supplemental Table 3.8.2. p53 and HRD-related features in HCC stratified by 199-Repair groups

	1 - Low	2 - High	p-value
n	**216**	**158**	
p53 status			
p53 mutant	45 (24.6)	138 (75.4)	<0.001
WT	171 (89.5)	20 (10.5)	
TP53 mutation status			
TP53 mutation	41 (36.6)	71 (63.4)	<0.001
WT	175 (66.8)	87 (33.2)	
HRD status			
HRD High	48 (36.1)	85 (63.9)	<0.001
HRD Low	153 (71.5)	61 (28.5)	
HBV status			
Negative	146 (67.3)	71 (32.7)	<0.001
Positive	66 (43.7)	85 (56.3)	

Supplemental Table 3.8.3. p53 and HRD-related features in HCC by Low repair subgroups

	L1	L2	L3	p-value
n	**86**	**56**	**74**	
p53 status				
p53 mutant	0 (0.0)	26 (57.8)	19 (42.2)	<0.001
WT	86 (50.3)	30 (17.5)	55 (32.2)	
TP53 mutation status				
TP53 mut	4 (9.7)	12 (29.3)	25 (61.0)	<0.001
WT	82 (46.9)	44 (25.1)	49 (28.0)	
HRD status				
HRD High	8 (16.7)	18 (37.5)	22 (45.8)	0.002
HRD Low	69 (45.1)	35 (22.9)	49 (32.0)	

CHAPTER 4: CONCLUSIONS

4.1 Summary

Primary liver cancer includes both hepatocellular carcinoma (HCC) and cholangiocarcinoma (CCA), despite them having distinct etiologic and genomic features. This work presents a more focused analysis on characterizing HCC to gain an understanding of biological heterogeneity. We have identified, using multiple genomic platforms, two molecular classes (CCA-like and Blast-like) separate from prototypical hepatocellular carcinoma tumors that are associated with distinct genomic patterns, clinical factors, and survival. These differences are likely due to several pathway interactions. Particularly, DNA repair pathways have been implicated in producing clinical variability in HCC outcomes[89,119-121].

DNA damage and repair pathways generate characteristic mutational signatures. These mutational signatures may result from endogenous mutagenic processes (e.g., DNA repair dysfunction) or exogenous processes (e.g., environmental carcinogens). Mutational signatures provide a method to directly measure repair deficiencies associated with genetic and epigenetic defects. We discovered that the types of mutational signatures and their prevalence varied across the HCC subclasses. CCA-like tumors had specific associations with DNA repair mutational signatures, such as mismatch repair (MMR), similar to CCAs. Whereas Blast-like tumors were associated with exposure-related signatures, such as liver toxin aflatoxin, parallel to HCC. Matching mutational signatures to DNA repair defects has the potential to uncover class-specific targeted therapies. Therefore, further evaluating RNA-based DNA repair signatures, in association with liver regeneration and HCC characteristics, is a promising strategy for establishing the etiology of these signatures, and their significance in HCC tumor

progression. This would allow discovery of biomarkers candidates for patients with intermediate and advanced HCC, for which treatment options are limited.

We identified two major repair classes in HCC characterized by low or high repair expression across all repair pathways similar to other studies[93-96]. However, the subset of samples with pronounced high repair expression across the board masked heterogeneity in the low repair samples. When we considered the low repair tumors separately, we identified three subgroups, L1, L2, L3. Overall, DNA repair expression varied across HCC tumors with a subset of high repair/regenerative tumors marked by substantial disrupted liver biology expression, p53 mutant expression, and poor outcomes. These results indicate that there is a dynamic balance between liver regeneration and DNA repair signals. Dysregulation in these processes promotes a pro-tumorigenic microenvironment that facilitates tumor progression and alters cells malignant potential causing further disrupted liver function. Interestingly, most of the CCA-like and Blast-like tumors associated with High repair/Regeneration, suggesting a link between repair dysfunction, mutational signatures, and outcomes. Our observations support the idea that repair signatures can be biomarkers for specific HCC characteristics and may bear clinical value as predictors of survival. The proposed method identifies additional biological heterogeneity beyond other two-class DNA repair schemas and the HCC tumor grading system alone.

4.2 Significance and Translational Implications

The prognostic ability of our molecular classes and DNA repair signatures poses the potential for them to be clinically relevant. Our current study evaluated this potential in HCC; however, future work should focus on evaluating DNA repair classes as prognostic and diagnostic signatures for predicting outcomes and patient classification Additional work to further delineate the important signaling pathway to better understand pathways required to drive malignant transformation of hepatocytes.

Although we are beginning to understand transdifferentiation and its role in HCC, much remains unknown. Our DNA repair signature strongly associates with the molecular subtype classes, with higher DNA repair correlating with more dedifferentiated classes Blast-like and CCA-like. CCA-like exhibited similar mutational patterns to CCA, reduced exposure-related mutational signatures compared to HCC, and upregulation of transdifferentiation pathways NOTCH, WNT and TGFβ. Mutational signatures provide a new way to examine exogenous and endogenous factors that have influenced cancer development. By examining mutational signatures and integrating them with our repair classes, we may reveal factors affecting prognosis in HCC tumors. Repair dysfunction in HCC tumors may interact with multiple signaling pathways to derive CCA-like and Blast-like tumors. Understanding these pathways could provide insight on progression and recurrence mechanisms, which is of significant importance since ~70% of all HCC cases recur even those with lower grade[7].

One major obstacle is high recurrence after surgical resection (up to 70% at 5 years, even at very early stage, single nodule ≤2cm)[1]. A second obstacle is ~90% of cases present in the setting of chronic liver damage, thereby limiting therapeutic options due to the patient's overall health condition[7]. Liver regeneration following chronic liver damage or surgical treatment has been implicated in stimulating tumor progression and micrometastases through activation of cellular signaling pathways, particularly DNA damage and repair pathways[225].

To extend our current work, our DNA repair signatures could be applied to HCC tumors treated with curative resection and/or systemic targeted agent, followed by evaluation of distinguishing features for groups and outcome measures. Due to the high proliferation associated with liver regeneration, tumors treated with surgery would be expected to display high repair/regenerative signatures with greater genomic instability. In addition, the past classification of viral and nonviral may not be sufficient to explain HCC biological and clinical heterogeneity. Similarly, other risk factors including non-alcoholic fatty liver disease (NAFLD)

and non-alcoholic steatohepatitis (NASH) could also affect HCC heterogeneity and may be related to differences in DNA damage and response.

Recent discoveries of molecular alterations underlying cancer has allowed for targeted treatments that are more selective and better against more aggressive malignancies[86,192,231,232]. Approximately 25% of HCCs have clinically actionable targets, yet none have been translated into the clinic[86]. This is complicated by the fact that the prevalence of most mutations is <10%[11]. We noted significant differences in prevalence of recurrent key drivers for HCC, namely *TP53* and *CTNNB1,* by molecular class and repair groups. Blast-like tumors exhibited an increased frequency of *TP53* mutations. Repair high and L3 samples showed an enrichment of TP53 mutations and p53 dysfunction, whereas the L1 and L2 repair tumors showed increased *CTNNB1* mutations. Indeed, our results highlight the importance of these drivers and pathway dysregulation in HCC heterogeneity. To date, targeted therapies to *TP53* and *CTNNB1* remain difficult but the repair signatures reveal other pathways involved that could guide biomarker and treatment development[86]. This is important because the main curative option remains surgical intervention. Additionally, as this disease typically is diagnosed at advanced stages, most patients present with unresectable tumors so they must undergo systemic therapies.

The DNA repair signatures have incorporated the unique properties of the liver, and with the incorporation of treatment data may elucidate biomarkers of response. Moreover, Immune checkpoint inhibitors have shown limited success in HCCs mediated in part by the immunosuppressive tumor microenvironment (TME)[233]. Combining ICIs with TKIs or VEGF inhibitors resulted in double response rates and survival benefits compared with single agents[234]. This could be advantageous in Blast-like and High repair samples where there are higher rates of chronic HBV infection, since HBV mediates immunosuppression and tumors tend to display increased genomic instability. Therapeutic agents that reverse immunosuppressive

nature of HCC tumors, alone or in combination with other therapies, are essential for improving outcomes for HCC patients.

Moreover, our results revealed several potential biomarkers, such as HRD, repair dysfunction, and p53 dysfunction, that are representative of HCC heterogeneity. Homologous recombination deficiency (HRD) is a promising biomarker for immunotherapy. HRD has a substantial influence on genomic stability and tumorigenesis. It can cause increased sensitivity to PARP inhibitors and platinum-based therapy. The high repair dysfunction and high HRD seen in our High repair and L3 tumors suggests targeting with PARP inhibitors or platinum-based therapies may improve outcomes for HCC patients with high HRD. High HRD score is associated with worse prognosis and has been shown to be predictive of clinical benefit with PARP inhibitors[120,208,209,235]. The development of DNA targeted therapies and immunotherapies for intermediate to high repair samples of utmost importance as they are higher grade samples for which there are limited treatment options, compared to low repair tumors. However, if future research shows that these signatures do not specifically predict response to these therapies, it is also possible that these signatures might guide surgical decision making, providing clinicians with insights as to the extent of resection needed for best prognosis.

4.3 Future Directions and Conclusions

There are questions that must be addressed to translate these findings to the clinic: First, can these signatures be adapted to a clinically-relevant platform applicable in diverse populations and second, how we can exploit genomic differences between the DNA repair classes for the development of targeted therapies or treatment decision making? Clinical biopsies often have limited material and require more accessible molecular platforms. Many molecular study cohorts contain are predominantly white participants; therefore, there is a great need for better representation of historically excluded populations. This is particularly important as these populations tend to present with more clinically aggressive phenotypes. The first step

to addressing health inequities in precision medicine is increasing availability of genomic data from diverse populations and more dedicated initiatives to studying the clinical implications of variants in these patient populations. However, it can be challenging to apply these signatures broadly to diverse populations because RNA may be limited, RNA sequencing costs are high, and access to care for diverse and medically underserved populations (i.e., lower diversity at tissue procurement sites).

One option is to transition the RNA repair signature to the NanoString platform, which is the basis for the FDA approved Prosigna assay for breast cancer[236,237], since it provides accurate quantification of gene expression for formalin fixed paraffin embedded (FFPE) samples, as well as fresh frozen tumors. The development and validation of signatures in NanoString could address knowledge gaps such as the ability of the signature to assess overall prognosis and determine correlations between DNA repair expression and response to therapy in diverse clinical settings.

Validation of these repair signatures in multiple cohorts and in association with therapeutic regimens would result in better classification and prognostication of HCC tumors. This would allow for the use of differential repair patterns to inform treatment options. Differing contributions of exposure-related and DNA repair mutational signatures to CCA-like and Blast-like tumors demonstrates clinical relevance that may also translate to other tumor types. The largest limitation of many current studies is the lack of available clinical and treatment data . As more therapeutic data and clinical annotations becomes available, these questions can be addressed.

REFERENCES

1. Llovet, J.M. *et al.* Hepatocellular carcinoma. *Nat Rev Dis Primers* **7**, 6 (2021).

2. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* **71**, 209-249 (2021).

3. Siegel, R.L., Miller, K.D., Fuchs, H.E. & Jemal, A. Cancer Statistics, 2021. *CA Cancer J Clin* **71**, 7-33 (2021).

4. Marrero, J.A. *et al.* Diagnosis, Staging, and Management of Hepatocellular Carcinoma: 2018 Practice Guidance by the American Association for the Study of Liver Diseases. *Hepatology* **68**, 723-750 (2018).

5. Valery, P.C. *et al.* Projections of primary liver cancer to 2030 in 30 countries worldwide. *Hepatology* **67**, 600-611 (2018).

6. Siegel, R.L., Miller, K.D., Fuchs, H.E. & Jemal, A. Cancer statistics, 2022. *CA Cancer J Clin* **72**, 7-33 (2022).

7. Yang, J.D. *et al.* A global view of hepatocellular carcinoma: trends, risk, prevention and management. *Nat Rev Gastroenterol Hepatol* **16**, 589-604 (2019).

8. Petrick, J.L., Kelly, S.P., Altekruse, S.F., McGlynn, K.A. & Rosenberg, P.S. Future of Hepatocellular Carcinoma Incidence in the United States Forecast Through 2030. *J Clin Oncol* **34**, 1787-94 (2016).

9. Shiels, M.S. & O'Brien, T.R. Recent Decline in Hepatocellular Carcinoma Rates in the United States. *Gastroenterology* **158**, 1503-1505 e2 (2020).

10. Wu, E.M. *et al.* Gender differences in hepatocellular cancer: disparities in nonalcoholic fatty liver disease/steatohepatitis and liver transplantation. *Hepatoma Res* **4**(2018).

11. Cancer Genome Atlas Research Network. Electronic address, w.b.e. & Cancer Genome Atlas Research, N. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell* **169**, 1327-1341 e23 (2017).

12. Schulze, K. *et al.* Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat Genet* **47**, 505-511 (2015).

13. Zamor, P.J., deLemos, A.S. & Russo, M.W. Viral hepatitis and hepatocellular carcinoma: etiology and management. *J Gastrointest Oncol* **8**, 229-242 (2017).

14. Hoshida, Y., Villanueva, A. & Llovet, J.M. Molecular profiling to predict hepatocellular carcinoma outcome. *Expert Rev Gastroenterol Hepatol* **3**, 101-3 (2009).

15. de Martel, C., Georges, D., Bray, F., Ferlay, J. & Clifford, G.M. Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis. *Lancet Glob Health* **8**, e180-e190 (2020).

16. Amaddeo, G. *et al.* Integration of tumour and viral genomic characterizations in HBV-related hepatocellular carcinomas. *Gut* **64**, 820-9 (2015).

17. Kew, M.C. Hepatitis B virus x protein in the pathogenesis of hepatitis B virus-induced hepatocellular carcinoma. *J Gastroenterol Hepatol* **26 Suppl 1**, 144-52 (2011).

18. Lim, H.C. & Gordan, J.D. Tumor hepatitis B virus RNA identifies a clinically and molecularly distinct subset of hepatocellular carcinoma. *PLoS Comput Biol* **17**, e1008699 (2021).

19. Medhat, A., Arzumanyan, A. & Feitelson, M.A. Hepatitis B x antigen (HBx) is an important therapeutic target in the pathogenesis of hepatocellular carcinoma. *Oncotarget* **12**, 2421-2433 (2021).

20. Peng, Z. *et al.* Integration of the hepatitis B virus X fragment in hepatocellular carcinoma and its effects on the expression of multiple molecules: a key to the cell cycle and apoptosis. *Int J Oncol* **26**, 467-73 (2005).

21. Ahn, S.M. *et al.* Genomic portrait of resectable hepatocellular carcinomas: implications of RB1 and FGF19 aberrations for patient stratification. *Hepatology* **60**, 1972-82 (2014).

22. Yao, S. *et al.* Differences in somatic mutation landscape of hepatocellular carcinoma in Asian American and European American populations. *Oncotarget* **7**, 40491-40499 (2016).

23. Maucort-Boulch, D., de Martel, C., Franceschi, S. & Plummer, M. Fraction and incidence of liver cancer attributable to hepatitis B and C viruses worldwide. *Int J Cancer* **142**, 2471-2477 (2018).

24. Axley, P., Ahmed, Z., Ravi, S. & Singal, A.K. Hepatitis C Virus and Hepatocellular Carcinoma: A Narrative Review. *J Clin Transl Hepatol* **6**, 79-84 (2018).

25. Mavilia, M.G. & Wu, G.Y. HBV-HCV Coinfection: Viral Interactions, Management, and Viral Reactivation. *J Clin Transl Hepatol* **6**, 296-305 (2018).

26. Liu, Y. & Wu, F. Global burden of aflatoxin-induced hepatocellular carcinoma: a risk assessment. *Environ Health Perspect* **118**, 818-24 (2010).

27. Strosnider, H. *et al.* Workgroup report: public health strategies for reducing aflatoxin exposure in developing countries. *Environ Health Perspect* **114**, 1898-903 (2006).

28. Zyoud, S.H. Global scientific trends on aflatoxin research during 1998-2017: a bibliometric and visualized study. *J Occup Med Toxicol* **14**, 27 (2019).

29. Fedeles, B.I. & Essigmann, J.M. Impact of DNA lesion repair, replication and formation on the mutational spectra of environmental carcinogens: Aflatoxin B1 as a case study. *DNA Repair (Amst)* **71**, 12-22 (2018).

30. Kew, M.C. Hepatocellular carcinoma: epidemiology and risk factors. *J Hepatocell Carcinoma* **1**, 115-25 (2014).

31. Huang, P.L. A comprehensive definition for metabolic syndrome. *Dis Model Mech* **2**, 231-7 (2009).

32. Moore, J.X., Chaudhary, N. & Akinyemiju, T. Metabolic Syndrome Prevalence by Race/Ethnicity and Sex in the United States, National Health and Nutrition Examination Survey, 1988-2012. *Prev Chronic Dis* **14**, E24 (2017).

33. Tsoulfas, G. Hepatocellular carcinoma and metabolic syndrome: The times are changing and so should we. *World J Gastroenterol* **25**, 3842-3848 (2019).

34. Anstee, Q.M., Reeves, H.L., Kotsiliti, E., Govaere, O. & Heikenwalder, M. From NASH to HCC: current concepts and future challenges. *Nat Rev Gastroenterol Hepatol* **16**, 411-428 (2019).

35. Younossi, Z.M. *et al.* Global epidemiology of nonalcoholic fatty liver disease-Meta-analytic assessment of prevalence, incidence, and outcomes. *Hepatology* **64**, 73-84 (2016).

36. Bellentani, S. The epidemiology of non-alcoholic fatty liver disease. *Liver Int* **37 Suppl 1**, 81-84 (2017).

37. Morgan, T.R., Mandayam, S. & Jamal, M.M. Alcohol and hepatocellular carcinoma. *Gastroenterology* **127**, S87-96 (2004).

38. Testino, G., Leone, S. & Borro, P. Alcohol and hepatocellular carcinoma: a review and a point of view. *World J Gastroenterol* **20**, 15943-54 (2014).

39. Llovet, J.M., Montal, R., Sia, D. & Finn, R.S. Molecular therapies and precision medicine for hepatocellular carcinoma. *Nat Rev Clin Oncol* **15**, 599-616 (2018).

40. Simon, T.G. *et al.* Association of Aspirin with Hepatocellular Carcinoma and Liver-Related Mortality. *N Engl J Med* **382**, 1018-1028 (2020).

41. Llovet, J.M. *et al.* Locoregional therapies in the era of molecular and immune treatments for hepatocellular carcinoma. *Nat Rev Gastroenterol Hepatol* **18**, 293-313 (2021).

42. European Association for the Study of the Liver. Electronic address, e.e.e. & European Association for the Study of the, L. EASL Clinical Practice Guidelines: Management of hepatocellular carcinoma. *J Hepatol* **69**, 182-236 (2018).

43. Kardashian, A. *et al.* Liver Transplantation Outcomes in a U.S. Multicenter Cohort of 789 Patients With Hepatocellular Carcinoma Presenting Beyond Milan Criteria. *Hepatology* **72**, 2014-2028 (2020).

44. Marisi, G. *et al.* Ten years of sorafenib in hepatocellular carcinoma: Are there any predictive and/or prognostic markers? *World J Gastroenterol* **24**, 4152-4163 (2018).

45. Finn, R.S. *et al.* Atezolizumab plus Bevacizumab in Unresectable Hepatocellular Carcinoma. *N Engl J Med* **382**, 1894-1905 (2020).

46. Llovet, J.M. *et al.* Sorafenib in advanced hepatocellular carcinoma. *N Engl J Med* **359**, 378-90 (2008).

47. Sacco, R. *et al.* Long-term results of sorafenib in advanced-stage hepatocellular carcinoma: what can we learn from routine clinical practice? *Expert Rev Anticancer Ther* **12**, 869-75 (2012).

48. Kudo, M. *et al.* Analysis of survival and objective response (OR) in patients with hepatocellular carcinoma in a phase III study of lenvatinib (REFLECT). *Journal of Clinical Oncology* **37**, 186-186 (2019).

49. Bruix, J. *et al.* Regorafenib for patients with hepatocellular carcinoma who progressed on sorafenib treatment (RESORCE): a randomised, double-blind, placebo-controlled, phase 3 trial. *Lancet* **389**, 56-66 (2017).

50. Abou-Alfa, G.K. *et al.* Cabozantinib in Patients with Advanced and Progressing Hepatocellular Carcinoma. *N Engl J Med* **379**, 54-63 (2018).

51. Zhu, A.X. *et al.* Ramucirumab after sorafenib in patients with advanced hepatocellular carcinoma and increased alpha-fetoprotein concentrations (REACH-2): a randomised, double-blind, placebo-controlled, phase 3 trial. *Lancet Oncol* **20**, 282-296 (2019).

52. El-Khoueiry, A.B. *et al.* Nivolumab in patients with advanced hepatocellular carcinoma (CheckMate 040): an open-label, non-comparative, phase 1/2 dose escalation and expansion trial. *Lancet* **389**, 2492-2502 (2017).

53. Finn, R.S. *et al.* Pembrolizumab As Second-Line Therapy in Patients With Advanced Hepatocellular Carcinoma in KEYNOTE-240: A Randomized, Double-Blind, Phase III Trial. *J Clin Oncol* **38**, 193-202 (2020).

54. Sia, D., Villanueva, A., Friedman, S.L. & Llovet, J.M. Liver Cancer Cell of Origin, Molecular Class, and Effects on Patient Prognosis. *Gastroenterology* **152**, 745-761 (2017).

55. Schulze, R.J., Schott, M.B., Casey, C.A., Tuma, P.L. & McNiven, M.A. The cell biology of the hepatocyte: A membrane trafficking machine. *J Cell Biol* **218**, 2096-2112 (2019).

56. Seehawer, M. *et al.* Necroptosis microenvironment directs lineage commitment in liver cancer. *Nature* **562**, 69-75 (2018).

57. Guichard, C. *et al.* Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. *Nat Genet* **44**, 694-8 (2012).

58. Llovet, J.M. *et al.* Hepatocellular carcinoma. *Nat Rev Dis Primers* **2**, 16018 (2016).

59. Chiang, D.Y. *et al.* Focal gains of VEGFA and molecular classification of hepatocellular carcinoma. *Cancer Res* **68**, 6779-88 (2008).

60. Hyman, D.M., Taylor, B.S. & Baselga, J. Implementing Genome-Driven Oncology. *Cell* **168**, 584-599 (2017).

61. Wang, B. *et al.* Null genotypes of GSTM1 and GSTT1 contribute to hepatocellular carcinoma risk: evidence from an updated meta-analysis. *J Hepatol* **53**, 508-18 (2010).

62. Romeo, S. *et al.* Genetic variation in PNPLA3 confers susceptibility to nonalcoholic fatty liver disease. *Nat Genet* **40**, 1461-5 (2008).

63. Nault, J.C., Ningarhari, M., Rebouissou, S. & Zucman-Rossi, J. The role of telomeres and telomerase in cirrhosis and liver cancer. *Nat Rev Gastroenterol Hepatol* **16**, 544-558 (2019).

64. Bisteau, X., Caldez, M.J. & Kaldis, P. The Complex Relationship between Liver Cancer and the Cell Cycle: A Story of Multiple Regulations. *Cancers (Basel)* **6**, 79-111 (2014).

65. Buch, S. *et al.* A genome-wide association study confirms PNPLA3 and identifies TM6SF2 and MBOAT7 as risk loci for alcohol-related cirrhosis. *Nat Genet* **47**, 1443-8 (2015).

66. Alexandrov, L.B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101 (2020).

67. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-21 (2013).

68. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J. & Stratton, M.R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* **3**, 246-59 (2013).

69. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47-54 (2016).

70. Letouze, E. *et al.* Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nat Commun* **8**, 1315 (2017).

71. Chawanthayatham, S. *et al.* Mutational spectra of aflatoxin B1 in vivo establish biomarkers of exposure for human hepatocellular carcinoma. *Proc Natl Acad Sci U S A* **114**, E3101-E3109 (2017).

72. Nik-Zainal, S. *et al.* The genome as a record of environmental exposure. *Mutagenesis* **30**, 763-70 (2015).

73. Jiang, T. *et al.* Predictors of Chemosensitivity in Triple Negative Breast Cancer: An Integrated Genomic Analysis. *PLoS medicine* **13**, e1002193-e1002193 (2016).

74. Davies, H. *et al.* HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat Med* **23**, 517-525 (2017).

75. Davies, H. *et al.* Whole-Genome Sequencing Reveals Breast Cancers with Mismatch Repair Deficiency. *Cancer Res* **77**, 4755-4762 (2017).

76. Maura, F. *et al.* Biological and prognostic impact of APOBEC-induced mutations in the spectrum of plasma cell dyscrasias and multiple myeloma cell lines. *Leukemia* **32**, 1044-1048 (2018).

77. Walker, B.A. *et al.* APOBEC family mutational signatures are associated with poor prognosis translocations in multiple myeloma. *Nat Commun* **6**, 6997 (2015).

78. Su, W.H., Jou, Y.S., Zhang, J.H., Ho, C.M. & Tai, D.I. Mutations in NOTCH1 and nucleotide excision repair genes are correlated with prognosis of hepatitis B virus-associated hepatocellular carcinoma. *J Cancer* **9**, 2678-2686 (2018).

79. Ceccaldi, R. *et al.* Homologous-recombination-deficient tumours are dependent on Polθ-mediated repair. *Nature* **518**, 258-262 (2015).

80. Ueda, S. *et al.* Evaluation of ERCC1 expression for cisplatin sensitivity in human hepatocellular carcinoma. *Ann Surg Oncol* **18**, 1204-11 (2011).

81. Chen, Y.J. *et al.* Sonic hedgehog signaling protects human hepatocellular carcinoma cells against ionizing radiation in an autocrine manner. *Int J Radiat Oncol Biol Phys* **80**, 851-9 (2011).

82. Yang, X.D. *et al.* PARP inhibitor Olaparib overcomes Sorafenib resistance through reshaping the pluripotent transcriptome in hepatocellular carcinoma. *Mol Cancer* **20**, 20 (2021).

83. Zhao, J. *et al.* Olaparib and enzalutamide synergistically suppress HCC progression via the AR-mediated miR-146a-5p/BRCA1 signaling. *FASEB J* **34**, 5877-5891 (2020).

84. Wang, C. *et al.* Rational combination therapy for hepatocellular carcinoma with PARP1 and DNA-PK inhibitors. *Proc Natl Acad Sci U S A* **117**, 26356-26365 (2020).

85. Zucman-Rossi, J., Villanueva, A., Nault, J.C. & Llovet, J.M. Genetic Landscape and Biomarkers of Hepatocellular Carcinoma. *Gastroenterology* **149**, 1226-1239 e4 (2015).

86. Rebouissou, S. & Nault, J.C. Advances in molecular classification and precision oncology in hepatocellular carcinoma. *J Hepatol* **72**, 215-229 (2020).

87. Sia, D. *et al.* Identification of an Immune-specific Class of Hepatocellular Carcinoma, Based on Molecular Features. *Gastroenterology* **153**, 812-826 (2017).

88. Hoshida, Y. *et al.* Integrative transcriptome analysis reveals common molecular subclasses of human hepatocellular carcinoma. *Cancer Res* **69**, 7385-92 (2009).

89. Lee, J.S. *et al.* A novel prognostic subtype of human hepatocellular carcinoma derived from hepatic progenitor cells. *Nat Med* **12**, 410-6 (2006).

90. Boyault, S. *et al.* Transcriptome classification of HCC is related to gene alterations and to new therapeutic targets. *Hepatology* **45**, 42-52 (2007).

91. Calderaro, J., Ziol, M., Paradis, V. & Zucman-Rossi, J. Molecular and histological correlations in liver cancer. *J Hepatol* **71**, 616-630 (2019).

92. Lachenmayer, A. *et al.* Wnt-pathway activation in two molecular classes of hepatocellular carcinoma and experimental modulation by sorafenib. *Clin Cancer Res* **18**, 4997-5007 (2012).

93. Lin, Z. *et al.* Prognostic value of DNA repair based stratification of hepatocellular carcinoma. *Sci Rep* **6**, 25999 (2016).

94. Oshi, M. *et al.* Enhanced DNA Repair Pathway is Associated with Cell Proliferation and Worse Survival in Hepatocellular Carcinoma (HCC). *Cancers (Basel)* **13**(2021).

95. Lin, P., Gao, R.Z., Wen, R., He, Y. & Yang, H. DNA Damage Repair Profiles Alteration Characterize a Hepatocellular Carcinoma Subtype With Unique Molecular and Clinicopathologic Features. *Front Immunol* **12**, 715460 (2021).

96. Chen, Y. *et al.* Identification of a DNA Damage Response and Repair-Related Gene-Pair Signature for Prognosis Stratification Analysis in Hepatocellular Carcinoma. *Front Pharmacol* **13**, 857060 (2022).

97. Li, L.Y., Guan, Y.D., Chen, X.S., Yang, J.M. & Cheng, Y. DNA Repair Pathways in Cancer Therapy and Resistance. *Front Pharmacol* **11**, 629266 (2020).

98. Torgovnick, A. & Schumacher, B. DNA repair mechanisms in cancer development and therapy. *Front Genet* **6**, 157 (2015).

99. Faraoni, I. & Graziani, G. Role of BRCA Mutations in Cancer Treatment with Poly(ADP-ribose) Polymerase (PARP) Inhibitors. *Cancers (Basel)* **10**(2018).

100. Barroso-Sousa, R. *et al.* Prevalence and mutational determinants of high tumor mutation burden in breast cancer. *Annals of Oncology* **31**, 387-394 (2020).

101. Jiang, Y. *et al.* Association between homologous recombination deficiency and tumor mutational burden in lung cancer. *Journal of Clinical Oncology* **38**, e21043-e21043 (2020).

102. Liu, Y.L. *et al.* BRCA Mutations, Homologous DNA Repair Deficiency, Tumor Mutational Burden, and Response to Immune Checkpoint Inhibition in Recurrent Ovarian Cancer. *JCO Precision Oncology*, 665-679 (2020).

103. Laurini, E. *et al.* Role of Rad51 and DNA repair in cancer: A molecular perspective. *Pharmacol Ther* **208**, 107492 (2020).

104. Hoppe, M.M. *et al.* Quantitative imaging of RAD51 expression as a marker of platinum resistance in ovarian cancer. *EMBO Mol Med* **13**, e13366 (2021).

105. Makino, E. *et al.* Targeting Rad51 as a strategy for the treatment of melanoma cells resistant to MAPK pathway inhibition. *Cell Death Dis* **11**, 581 (2020).

106. Matos-Rodrigues, G., Guirouilh-Barbat, J., Martini, E. & Lopez, B.S. Homologous recombination, cancer and the 'RAD51 paradox'. *NAR Cancer* **3**, zcab016 (2021).

107. Lu, Y., Chu, A., Turker, M.S. & Glazer, P.M. Hypoxia-induced epigenetic regulation and silencing of the BRCA1 promoter. *Mol Cell Biol* **31**, 3339-50 (2011).

108. Bindra, R.S. *et al.* Down-regulation of Rad51 and decreased homologous recombination in hypoxic cancer cells. *Mol Cell Biol* **24**, 8504-18 (2004).

109. Yang, S.F. *et al.* Involvement of DNA damage response pathways in hepatocellular carcinoma. *Biomed Res Int* **2014**, 153867 (2014).

110. Oliver, T.G. *et al.* Chronic cisplatin treatment promotes enhanced damage repair and tumor progression in a mouse model of lung cancer. *Genes Dev* **24**, 837-52 (2010).

111. Wang, L.E. *et al.* DNA repair capacity in peripheral lymphocytes predicts survival of patients with non-small-cell lung cancer treated with first-line platinum-based chemotherapy. *J Clin Oncol* **29**, 4121-8 (2011).

112. Neboori, H.J. *et al.* Low p53 binding protein 1 (53BP1) expression is associated with increased local recurrence in breast cancer patients treated with breast-conserving surgery and radiotherapy. *Int J Radiat Oncol Biol Phys* **83**, e677-83 (2012).

113. Lee, J.M., Ledermann, J.A. & Kohn, E.C. PARP Inhibitors for BRCA1/2 mutation-associated and BRCA-like malignancies. *Ann Oncol* **25**, 32-40 (2014).

114. Dziadkowiec, K.N., Gasiorowska, E., Nowak-Markwitz, E. & Jankowska, A. PARP inhibitors: review of mechanisms of action and BRCA1/2 mutation targeting. *Prz Menopauzalny* **15**, 215-219 (2016).

115. Robson, M.E. *et al.* OlympiAD final overall survival and tolerability results: Olaparib versus chemotherapy treatment of physician's choice in patients with a germline BRCA mutation and HER2-negative metastatic breast cancer. *Ann Oncol* **30**, 558-566 (2019).

116. Fukumoto, T. *et al.* N(6)-Methylation of Adenosine of FZD10 mRNA Contributes to PARP Inhibitor Resistance. *Cancer Res* **79**, 2812-2820 (2019).

117. Watson, Z.L. *et al.* Histone methyltransferases EHMT1 and EHMT2 (GLP/G9A) maintain PARP inhibitor resistance in high-grade serous ovarian carcinoma. *Clin Epigenetics* **11**, 165 (2019).

118. Le, D.T. *et al.* Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* **357**, 409-413 (2017).

119. Zhu, W. *et al.* Identification of DNA repair-related genes predicting pathogenesis and prognosis for liver cancer. *Cancer Cell Int* **21**, 81 (2021).

120. Chen, J. *et al.* The prognostic value of HRD mutations in liver cancer. *Journal of Clinical Oncology* **38**, e13546-e13546 (2020).

121. Barash, H. *et al.* Accelerated carcinogenesis following liver regeneration is associated with chronic inflammation-induced double-strand DNA breaks. *Proc Natl Acad Sci U S A* **107**, 2207-12 (2010).

122. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **68**, 394-424 (2018).

123. Siegel, R.L., Miller, K.D. & Jemal, A. Cancer statistics, 2019. *CA Cancer J Clin* **69**, 7-34 (2019).

124. Yu, M.C., Yuan, J.M. & Lu, S.C. Alcohol, cofactors and the genetics of hepatocellular carcinoma. *J Gastroenterol Hepatol* **23 Suppl 1**, S92-7 (2008).

125. Charbel, H. & Al-Kawas, F.H. Cholangiocarcinoma: epidemiology, risk factors, pathogenesis, and diagnosis. *Curr Gastroenterol Rep* **13**, 182-7 (2011).

126. Khan, S.A., Thomas, H.C., Davidson, B.R. & Taylor-Robinson, S.D. Cholangiocarcinoma. *Lancet* **366**, 1303-14 (2005).

127. TCGA. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell* **169**, 1327-1341 e23 (2017).

128. Farshidfar, F. *et al.* Integrative Genomic Analysis of Cholangiocarcinoma Identifies Distinct IDH-Mutant Molecular Profiles. *Cell Rep* **18**, 2780-2794 (2017).

129. Woo, H.G. *et al.* Identification of a cholangiocarcinoma-like gene expression trait in hepatocellular carcinoma. *Cancer Res* **70**, 3034-41 (2010).

130. Chaisaingmongkol, J. *et al.* Common Molecular Subtypes Among Asian Hepatocellular Carcinoma and Cholangiocarcinoma. *Cancer Cell* **32**, 57-70 e3 (2017).

131. Yang, L. *et al.* A single-cell transcriptomic analysis reveals precise pathways and regulatory mechanisms underlying hepatoblast differentiation. *Hepatology* **66**, 1387-1401 (2017).

132. Selitsky, S.R., Marron, D., Mose, L.E., Parker, J.S. & Dittmer, D.P. Epstein-Barr Virus-Positive Cancers Show Altered B-Cell Clonality. *mSystems* **3**(2018).

133. Hu, H. *et al.* Long-Term Expansion of Functional Mouse and Human Hepatocytes as 3D Organoids. *Cell* **175**, 1591-1606 e19 (2018).

134. Wilkerson, M.D. & Hayes, D.N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572-1573 (2010).

135. Ben-Porath, I. *et al.* An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. *Nat Genet* **40**, 499-507 (2008).

136. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).

137. Mose, L.E., Perou, C.M. & Parker, J.S. Improved Indel Detection in DNA and RNA via Realignment with ABRA2. *Bioinformatics* (2019).

138. Tischler, G. & Lenonard, S. biobambam: tools for read pair collation based algorithms on BAM files. *Source Code for Biology and Medicine* **9**(2014).

139. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* **2**, 401-4 (2012).

140. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* **6**, pl1 (2013).

141. Durinck, S., Spellman, P.T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* **4**, 1184-91 (2009).

142. Walter, V., Du, Y., Danilova, L., Hayward, M.C. & Hayes, D.N. MVisAGe Identifies Concordant and Discordant Genomic Alterations of Driver Genes in Squamous Tumors. *Cancer Res* **78**, 3375-3385 (2018).

143. Weigman, V.J. *et al.* Basal-like Breast cancer DNA copy number losses identify genes involved in genomic instability, response to therapy, and patient survival. *Breast Cancer Res Treat* **133**, 865-80 (2012).

144. Gehring, J.S., Fischer, B., Lawrence, M. & Huber, W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics* **31**, 3673-5 (2015).

145. Wardell, C.P. barplot3d. (2019).

146. Dabney, A.R. Classification of microarrays to nearest centroids. *Bioinformatics* **21**, 4148-54 (2005).

147. Kassambara, A., Kosinski, M. & Biecek, P. survminer: Drawing Survival Curves using 'ggplot2'. *R package version 0.4.9* (2021).

148. Liu, J. *et al.* An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* **173**, 400-416 e11 (2018).

149. Roessler, S. *et al.* A unique metastasis gene signature enables prediction of tumor relapse in early-stage hepatocellular carcinoma patients. *Cancer Res* **70**, 10202-12 (2010).

150. Andersen, J.B. *et al.* Genomic and genetic characterization of cholangiocarcinoma identifies therapeutic targets for tyrosine kinase inhibitors. *Gastroenterology* **142**, 1021-1031 e15 (2012).

151. Xue, R. *et al.* Genomic and Transcriptomic Profiling of Combined Hepatocellular and Intrahepatic Cholangiocarcinoma Reveals Distinct Molecular Subtypes. *Cancer Cell* (2019).

152. Chaudhari, P., Tian, L., Deshmukh, A. & Jang, Y.Y. Expression kinetics of hepatic progenitor markers in cellular models of human liver development recapitulating hepatocyte and biliary cell fate commitment. *Exp Biol Med (Maywood)* **241**, 1653-62 (2016).

153. Malta, T.M. *et al.* Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation. *Cell* **173**, 338-354 e15 (2018).

154. Bindea, G. *et al.* Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity* **39**, 782-95 (2013).

155. Mootha, V.K. *et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* **34**, 267-73 (2003).

156. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).

157. Limaye, P.B., Bowen, W.C., Orr, A., Apte, U.M. & Michalopoulos, G.K. Expression of hepatocytic- and biliary-specific transcription factors in regenerating bile ducts during hepatocyte-to-biliary epithelial cell transdifferentiation. *Comp Hepatol* **9**, 9 (2010).

158. Schaub, J.R. *et al.* De novo formation of the biliary system by TGFbeta-mediated hepatocyte transdifferentiation. *Nature* **557**, 247-251 (2018).

159. Sekiya, S. & Suzuki, A. Intrahepatic cholangiocarcinoma can arise from Notch-mediated conversion of hepatocytes. *J Clin Invest* **122**, 3914-8 (2012).

160. Terada, M. *et al.* Kupffer cells induce Notch-mediated hepatocyte conversion in a common mouse model of intrahepatic cholangiocarcinoma. *Sci Rep* **6**, 34691 (2016).

161. Wu, W.R. *et al.* Notch1 is overexpressed in human intrahepatic cholangiocarcinoma and is associated with its proliferation, invasiveness and sensitivity to 5-fluorouracil in vitro. *Oncol Rep* **31**, 2515-24 (2014).

162. Okabe, H. *et al.* Wnt signaling regulates hepatobiliary repair following cholestatic liver injury in mice. *Hepatology* **64**, 1652-1666 (2016).

163. Huntzicker, E.G. *et al.* Differential effects of targeting Notch receptors in a mouse model of liver cancer. *Hepatology* **61**, 942-52 (2015).

164. Jopling, C.L., Yi, M., Lancaster, A.M., Lemon, S.M. & Sarnow, P. Modulation of hepatitis C virus RNA abundance by a liver-specific MicroRNA. *Science* **309**, 1577-81 (2005).

165. Xue, R. *et al.* Variable Intra-Tumor Genomic Heterogeneity of Multiple Lesions in Patients With Hepatocellular Carcinoma. *Gastroenterology* **150**, 998-1008 (2016).

166. Ozturk, M. p53 mutation in hepatocellular carcinoma after aflatoxin exposure. *Lancet* **338**, 1356-9 (1991).

167. Polaris Observatory, C. Global prevalence, treatment, and prevention of hepatitis B virus infection in 2016: a modelling study. *Lancet Gastroenterol Hepatol* **3**, 383-403 (2018).

168. Hou, X.J. *et al.* Immune response involved in liver damage and the activation of hepatic progenitor cells during liver tumorigenesis. *Cell Immunol* **326**, 52-59 (2018).

169. Oh, I.S. & Park, S.H. Immune-mediated Liver Injury in Hepatitis B Virus Infection. *Immune Netw* **15**, 191-8 (2015).

170. Shin, E.C., Sung, P.S. & Park, S.H. Immune responses and immunopathology in acute and chronic viral hepatitis. *Nat Rev Immunol* **16**, 509-23 (2016).

171. Kim, H.S., Park, E.J. & Lee, C.W. Implication of hepatocyte dedifferentiation in pathogenesis and treatment of hepatocellular carcinoma. *Precision and Future Medicine* **3**, 37-42 (2019).

172. Hirsch, T.Z. *et al.* BAP1 mutations define a homogeneous subgroup of hepatocellular carcinoma with fibrolamellar-like features and activated PKA. *J Hepatol* **72**, 924-936 (2020).

173. Wardell, C.P. *et al.* Genomic characterization of biliary tract cancers identifies driver genes and predisposing mutations. *J Hepatol* **68**, 959-969 (2018).

174. Artegiani, B. *et al.* Probing the Tumor Suppressor Function of BAP1 in CRISPR-Engineered Human Liver Organoids. *Cell Stem Cell* **24**, 927-943 e6 (2019).

175. Xue, R. *et al.* Genomic and Transcriptomic Profiling of Combined Hepatocellular and Intrahepatic Cholangiocarcinoma Reveals Distinct Molecular Subtypes. *Cancer Cell* **35**, 932-947 e8 (2019).

176. Zheng, Q. *et al.* Immune signature-based hepatocellular carcinoma subtypes may provide novel insights into therapy and prognosis predictions. *Cancer Cell Int* **21**, 330 (2021).

177. Zhuang, W. *et al.* An immunogenomic signature for molecular classification in hepatocellular carcinoma. *Mol Ther Nucleic Acids* **25**, 105-115 (2021).

178. Damrauer, J.S. *et al.* Genomic characterization of rare molecular subclasses of hepatocellular carcinoma. *Commun Biol* **4**, 1150 (2021).

179. Weglarz, T.C. & Sandgren, E.P. Timing of hepatocyte entry into DNA synthesis after partial hepatectomy is cell autonomous. *Proc Natl Acad Sci U S A* **97**, 12595-600 (2000).

180. Caldez, M.J., Bjorklund, M. & Kaldis, P. Cell cycle regulation in NAFLD: when imbalanced metabolism limits cell division. *Hepatol Int* **14**, 463-474 (2020).

181. Greenbaum, L.E. Cell cycle regulation and hepatocarcinogenesis. *Cancer Biol Ther* **3**, 1200-7 (2004).

182. Yang, S.Q., Lin, H.Z., Mandal, A.K., Huang, J. & Diehl, A.M. Disrupted signaling and inhibited regeneration in obese mice with fatty livers: implications for nonalcoholic fatty liver disease pathophysiology. *Hepatology* **34**, 694-706 (2001).

183. Zhao, G., Nakano, K., Chijiiwa, K., Ueda, J. & Tanaka, M. Inhibited activities in CCAAT/enhancer-binding protein, activating protein-1 and cyclins after hepatectomy in rats with thioacetamide-induced liver cirrhosis. *Biochem Biophys Res Commun* **292**, 474-81 (2002).

184. Fausto, N., Campbell, J.S. & Riehle, K.J. Liver regeneration. *Hepatology* **43**, S45-53 (2006).

185. Furchtgott, L.A., Chow, C.C. & Periwal, V. A model of liver regeneration. *Biophys J* **96**, 3926-35 (2009).

186. Dewhurst, M.R. *et al.* Loss of hepatocyte cell division leads to liver inflammation and fibrosis. *PLoS Genet* **16**, e1009084 (2020).

187. Miyaoka, Y. & Miyajima, A. To divide or not to divide: revisiting liver regeneration. *Cell Div* **8**, 8 (2013).

188. Fortini, P., Ferretti, C. & Dogliotti, E. The response to DNA damage during differentiation: pathways and consequences. *Mutat Res* **743-744**, 160-168 (2013).

189. Chung, Y.L. & Wu, M.L. Promyelocytic leukaemia protein links DNA damage response and repair to hepatitis B virus-related hepatocarcinogenesis. *J Pathol* **230**, 377-87 (2013).

190. Bayram, S., Akkiz, H., Bekar, A., Akgollu, E. & Yildirim, S. The significance of Exonuclease 1 K589E polymorphism on hepatocellular carcinoma susceptibility in the Turkish population: a case-control study. *Mol Biol Rep* **39**, 5943-51 (2012).

191. Lees-Miller, S.P. & Meek, K. Repair of DNA double strand breaks by non-homologous end joining. *Biochimie* **85**, 1161-73 (2003).

192. Okamura, R. *et al.* Comprehensive genomic landscape and precision therapeutic approach in biliary tract cancers. *Int J Cancer* **148**, 702-712 (2021).

193. Rao, S. *et al.* Quantification and expert evaluation of evidence for chemopredictive biomarkers to personalize cancer treatment. *Oncotarget* **8**, 37923-37934 (2017).

194. Ellrott, K. *et al.* Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst* **6**, 271-281 e7 (2018).

195. Hoadley, K.A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291-304 e6 (2018).

196. Gao, Y. *et al.* A neomorphic cancer cell-specific role of MAGE-A4 in trans-lesion synthesis. *Nature Communications* **7**, 12105 (2016).

197. Gao, Y. *et al.* The Cancer/Testes (CT) Antigen HORMAD1 promotes Homologous Recombinational DNA Repair and Radioresistance in Lung adenocarcinoma cells. *Scientific Reports* **8**, 15304 (2018).

198. Scrucca, L., Fop, M., Murphy, T.B. & Raftery, A.E. mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *R J* **8**, 289-317 (2016).

199. Selitsky, S.R. *et al.* Virus expression detection reveals RNA-sequencing contamination in TCGA. *BMC Genomics* **21**, 79 (2020).

200. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847-9 (2016).

201. R Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing Vienna, Austria* (2021).

202. Colak, D. *et al.* RNA-Seq transcriptome profiling in three liver regeneration models in rats: comparative analysis of partial hepatectomy, ALLPS, and PVL. *Sci Rep* **10**, 5213 (2020).

203. Abkevich, V. *et al.* Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *British journal of cancer* **107**, 1776-1782 (2012).

204. Popova, T. *et al.* Ploidy and Large-Scale Genomic Instability Consistently Identify Basal-like Breast Carcinomas with BRCA1/2 Inactivation. *Cancer Research* **72**, 5454 (2012).

205. Birkbak, N.J. *et al.* Telomeric Allelic Imbalance Indicates Defective DNA Repair and Sensitivity to DNA-Damaging Agents. *Cancer Discovery* **2**, 366 (2012).

206. Knijnenburg, T.A. *et al.* Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. *Cell reports* **23**, 239-254.e6 (2018).

207. Telli, M.L. *et al.* Evaluation of homologous recombination deficiency (HRD) status with pathological response to carboplatin +/- veliparib in BrighTNess, a randomized phase 3 study in early stage TNBC. *Journal of Clinical Oncology* **36**, 519-519 (2018).

208. Takaya, H., Nakai, H., Takamatsu, S., Mandai, M. & Matsumura, N. Homologous recombination deficiency status-based classification of high-grade serous ovarian carcinoma. *Sci Rep* **10**, 2757 (2020).

209. Wen, H. *et al.* Homologous recombination deficiency in diverse cancer types and its correlation with platinum chemotherapy efficiency in ovarian cancer. *BMC Cancer* **22**, 550 (2022).

210. Edge, S.B. & American Joint Committee on Cancer. *AJCC cancer staging manual*, xiv, 648 p. (Springer, New York ; London, 2010).

211. Greene, F.L., American Joint Committee on Cancer., American Cancer Society. & Sadie Jenkins Harmon Collection. *AJCC cancer staging manual*, xiv, 421 pages (Springer-Verlag, New York, 2002).

212. Fleming, I.D. *et al. AJCC cancer staging manual*, xv, 294 pages (Lippincott-Raven, Philadelphia, 1997).

213. Beahrs, O.H. *et al. AJCC cancer staging manual*, xv, 293 (Philadelphia, Lippincott-Raven, 1977).

214. Troester, M.A. *et al.* Gene expression patterns associated with p53 status in breast cancer. *BMC Cancer* **6**, 276 (2006).

215. Yang, W.X., Pan, Y.Y. & You, C.G. CDK1, CCNB1, CDC20, BUB1, MAD2L1, MCM3, BUB1B, MCM2, and RFC4 May Be Potential Therapeutic Targets for Hepatocellular Carcinoma Using Integrated Bioinformatic Analysis. *Biomed Res Int* **2019**, 1245072 (2019).

216. Zhu, L.J., Pan, Y., Chen, X.Y. & Hou, P.F. BUB1 promotes proliferation of liver cancer cells by activating SMAD2 phosphorylation. *Oncol Lett* **19**, 3506-3512 (2020).

217. Qiu, J. *et al.* BUB1B promotes hepatocellular carcinoma progression via activation of the mTORC1 signaling pathway. *Cancer Med* **9**, 8159-8172 (2020).

218. Zhuang, L., Zhang, Y., Meng, Z. & Yang, Z. Oncogenic Roles of RAD51AP1 in Tumor Tissues Related to Overall Survival and Disease-Free Survival in Hepatocellular Carcinoma. *Cancer Control* **27**, 1073274820977149 (2020).

219. Duan, X. *et al.* The effect of the TP53 and RB1 mutations on the survival of hepatocellular carcinoma patients with different racial backgrounds. *J Gastrointest Oncol* **12**, 1786-1796 (2021).

220. Zong, B. *et al.* HORMAD1 promotes docetaxel resistance in triple negative breast cancer by enhancing DNA damage tolerance. *Oncol Rep* **46**(2021).

221. Shi, J.H. & Line, P.D. Effect of liver regeneration on malignant hepatic tumors. *World J Gastroenterol* **20**, 16167-77 (2014).

222. Li, H. & Zhang, L. Liver regeneration microenvironment of hepatocellular carcinoma for prevention and therapy. *Oncotarget* **8**, 1805-1813 (2017).

223. Zhong, C. *et al.* Serum and tissue vascular endothelial growth factor predicts prognosis in hepatocellular carcinoma patients after partial liver resection. *Hepatogastroenterology* **59**, 93-7 (2012).

224. Ding, T. *et al.* Endothelium-coated tumor clusters are associated with poor prognosis and micrometastasis of hepatocellular carcinoma after resection. *Cancer* **117**, 4878-89 (2011).

225. Michalopoulos, G.K. & Bhushan, B. Liver regeneration: biological and pathological mechanisms and implications. *Nat Rev Gastroenterol Hepatol* **18**, 40-55 (2021).

226. Krizhanovsky, V. *et al.* Senescence of activated stellate cells limits liver fibrosis. *Cell* **134**, 657-67 (2008).

227. Ha, G.H. & Breuer, E.K. Mitotic Kinases and p53 Signaling. *Biochem Res Int* **2012**, 195903 (2012).

228. Kay, J., Thadhani, E., Samson, L. & Engelward, B. Inflammation-induced DNA damage, mutations and cancer. *DNA Repair (Amst)* **83**, 102673 (2019).

229. Mei, J. *et al.* BRCA1 Is a Novel Prognostic Indicator and Associates with Immune Cell Infiltration in Hepatocellular Carcinoma. *DNA Cell Biol* **39**, 1838-1849 (2020).

230. Hou, Y. *et al.* Cyclin B1 acts as a tumor microenvironment-related cancer promoter and prognostic biomarker in hepatocellular carcinoma. *J Int Med Res* **49**, 3000605211016265 (2021).

231. Huang, R. & Zhou, P.K. DNA damage repair: historical perspectives, mechanistic pathways and clinical translation for targeted cancer therapy. *Signal Transduct Target Ther* **6**, 254 (2021).

232. Sullivan, K.M., Kenerson, H.L., Pillarisetty, V.G., Riehle, K.J. & Yeung, R.S. Precision oncology in liver cancer. *Ann Transl Med* **6**, 285 (2018).

233. Liu, H.T. *et al.* Immune Checkpoint Inhibitors in Hepatocellular Carcinoma: Current Progresses and Challenges. *Front Oncol* **11**, 737497 (2021).

234. Finn, R.S. *et al.* Phase Ib Study of Lenvatinib Plus Pembrolizumab in Patients With Unresectable Hepatocellular Carcinoma. *J Clin Oncol* **38**, 2960-2970 (2020).

235. Knijnenburg, T.A. *et al.* Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. *Cell Rep* **23**, 239-254 e6 (2018).

236. Parker, J.S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* **27**, 1160-7 (2009).

237. Prat, A. *et al.* Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res* **12**, R68 (2010).