

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Utilizing short sequences missing from the genome to identify gene regulatory element mutations associated with liver cancer

Permalink

<https://escholarship.org/uc/item/8fz1q6z2>

Author

Sims, Jasmine

Publication Date

2024

Peer reviewed|Thesis/dissertation

Utilizing short sequences missing from the genome to identify gene regulatory element mutations associated with liver cancer

by
Jasmine Sims

DISSERTATION
Submitted in partial satisfaction of the requirements for degree of
DOCTOR OF PHILOSOPHY


in

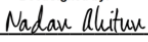
Biochemistry and Molecular Biology

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved:

DocuSigned by:
 LUKE GILBERT
71F73C69F83C48B... _____ Chair

DocuSigned by:
 Nadav Ahituv
D22A989F8E594F4... _____ Nadav Ahituv

DocuSigned by:
 Thomas Michael McManus
B47E685F33D1455... _____ Thomas Michael McManus

Committee Members

Dedication

I dedicate my thesis work to my Dad, Joseph Sims, and my grandmother, Ella Mae Sims. My dad pushed and inspired me to become the scientist that I am today; he gave me tenacity. My grandmother shaped me into the person that I am today, she gave me courage and determination despite my fears. I'd also like to thank my family for their support.

Acknowledgements

First, I'd like to thank God. Next, I would like to thank my thesis advisors, Dr. Nadav Ahituv and Franklin Huang, as well as my thesis committee members, Luke Gilbert and Michael McManus for their support over the years. I couldn't have asked for a better committee; they supported my research ideas along the way and challenged me in my scientific thinking in ways that improved both my scientific rigor as well as my project progression. They were very good about providing resources when asked and even allowed me to attend their lab meetings to learn more about science at a time when I wasn't very confident with my science. Thank you to Franklin Huang, Michael McManus and Luke Gilbert for all of your questions, and keeping me engaged and enthralled with science. Franklin Huang and Michael McManus, thank you both for opening up your labs to me and allowing me to be part of your lab family, it really meant a lot. Thank you Nadav Ahituv for teaching me how to write grants and going through twelve drafts back and forth with this article. Writing was not my strong suit, but he made me practice it and now I'm a much better writer and even go out of my way to help my scientific friends with their grants because of his influence.

Thank you to all members of the Ahituv, Huang, and McManus labs, especially Hanbing Song, who was an excellent and extremely helpful bioinformatics mentor. I sincerely enjoyed studying prostate cancer and learning single cell RNA-seq data analysis with him. He made comparing data findings fun and helped me level up on my coding. Thank you also to Jingjing Zhao, she helped me a lot when I first started coding, and never judged me for the silly coding questions

I'd ask her. I'd also like to thank Weiyu Li and Chengyu Deng, I sincerely learned a lot from you both about wet lab techniques, as well as experimental setup.

I'd also like to thank my previous thesis committee advisors from my master's, Mark Chan, Blake Riggs and Carmen Domingo. Mark was an amazing PI, that taught me to be curious in science and was a great mentor who cared about my research as well as my well-being. He was more than a mentor, he was a friend. Dr. Riggs was like a second father to me, and I am so thankful that I had the opportunity to learn from him, both in science and in life. I'd also like to thank Pleuni Pennings, she planted the seed in my coding ability through a bioinformatics class, and that seed grew into the bioinformatics person that I am today. I'd also like to thank Dr. Frank Bayliss, he saw something in me that I didn't have the confidence yet to see in myself, and he worked with me and helped me get my first real scholarship for science, and he treated me like a daughter. Thank you for believing in me.

Finally, I'd like to thank my friends, both from UCSF and beyond. Thank you to Cecelia Brown, Chinomnso Okorie, Jessica Bolivar, Sita Chandrasekaran, Tobi Thomas, Dina Buitrago, Christina Stephens, Jayson Davidson, Oley Drammeh, Muryam Gourdet, Sydney Williams, Imani Warren, Lyra Dimaposoc, Fred Santana, Nadia Ayad, Chase Webb, Tianna Grant, D'Anne Duncan, Sean Ganther, and so many others who've inspired me on this PhD journey and in life.

Contributions

Chapter 2 of this dissertation is an article in preparation, first authored by Jasmine Sims. This project has contributions from: Ofer Yizhar Barnea, Chengyu Deng, Weiyu Li, Jingjing Zhao, Ryder Easterlin, Ilias Georgakopoulos-Soares, Franklin Huang, Martin Hemberg, and Nadav Ahituv.

Utilizing short sequences missing from the genome to identify gene
regulatory element mutations associated with liver cancer

Jasmine Sims

Abstract

Liver cancer is diagnosed yearly in over 800,000 people worldwide. While several gene coding mutations have been found to be associated with this cancer, we lack an understanding of gene regulatory driver mutations that could also lead to liver cancer. Using short sequences, 16 base pairs in length, termed neomers, that are generally absent from the human genome but appear in liver cancer, we identify numerous potential causative gene regulatory mutations in promoters, potential enhancers and 3'UTRs. We identify noncoding neomers that are enriched in numerous patients and are thought to regulate cancer-associated genes. We also characterize specific one kilobase regions in the human genome that are enriched for noncoding neomers that reside near many relevant liver cancer genes. Combined, our study identifies novel mutations in gene regulatory elements that could be key regulators of liver cancer, providing a basis for potential downstream diagnosis and therapeutics.

Table of Contents

CHAPTER 1: INTRODUCTION	1
1.1 SYMPTOMS, CAUSES, TREATMENT, AND STAGES OF LIVER CANCER	1
1.2 RECURRENT CODING MUTATIONS IN LIVER CANCER.....	1
1.3 MUTATIONS IN GENE REGULATORY ELEMENTS IN LIVER CANCER	2
1.4 SUPER ENHANCER LANDSCAPES DIFFER BETWEEN LIVER CANCER AND NORMAL CELLS ..	4
1.5 UTILIZING NEOMERS TO IDENTIFY NONCODING RECURRENT MUTATIONS IN LIVER CANCER	5
CHAPTER 2: UTILIZING SHORT SEQUENCES MISSING FROM THE GENOME TO IDENTIFY GENE REGULATORY ELEMENT MUTATIONS ASSOCIATED WITH LIVER CANCER	13
2.1 SUMMARY	14
2.2 INTRODUCTION	15
2.3 RESULTS	18
2.3.2 ANNOTATION OF LIVER CANCER ASSOCIATED NEOMERS IN ENHANCERS	21
2.3.3 LIVER CANCER ASSOCIATED NEOMERS IN THE 3' UNTRANSLATED REGION.....	24
2.3.4 NEOMERS RECURRENT IN LIVER CANCER PATIENTS	27
2.3.5 NEOMERS RECURRENT IN 1KB REGIONS	28
DISCUSSION	30
METHODS.....	32
GENOMIC FEATURE ANNOTATION.....	32
NEOMER GENE TARGET LOCI IDENTIFICATION OF NONCODING REGIONS IN LIVER CANCER	32
GENE ONTOLOGY.....	33

PATIENT RECURRENCE ANALYSIS.....	33
1KB BIN ENRICHMENT ANALYSIS.....	33
REFERENCES	35

List of Figures

Figure 1.1 Stages of degradation of Liver function leading to Hepatocellular Carcinoma (HCC) and the respective treatment options for each stage of liver damage ⁴	2
Figure 1.2 Diagram describing how neomers, DNA sequences absent from the human genome of healthy patients, can be used as a tool to dissect somatic mutations that may be associated with cancer.	6
Figure 1.3 Schematic overview of neomer characterization in candidate cis-regulatory elements.....	8
Figure 2.1 Characterization of promoter neomers in liver cancer patients.....	20
Figure 2.2 Liver cancer-associated neomers in potential enhancers.....	23
Figure 2.3 Characterization of liver cancer-associated neomers in 3' UTRs.....	26

List of Tables

Table 2.1 The top 10 recurrent neomers in liver cancer patients.	28
Table 2.2 The top 10 1kb genomic neomer containing regions.	29

Chapter 1: Introduction

1.1 Symptoms, causes, treatment, and stages of liver cancer

Liver cancer is the seventh most common form of cancer¹, with over 800,000 people worldwide estimated to be diagnosed for liver cancer each year, and 700,000 passing away from liver cancer per year¹. Genetic factors associated with liver cancer include metabolic syndrome, diabetes, and male gender, with males being more prone than females². Normally, symptoms do not appear in the early stages of liver cancer. However, later, symptoms include weight loss, upper abdominal pain, or yellowing of the skin (jaundice)³. Based on the patient's tumor stage, liver function, and performance status, a variety of options are available for liver cancer treatment options, including liver transplantation, surgical resection, percutaneous ablation, chemotherapy radiation and immunotherapy (**Figure 1.1**)⁴. While the liver can cope with small losses in liver function and can regenerate or reversibly heal itself upon proper lifestyle changes, once the liver reaches fibrosis, with liver function between 20-30%, the stages of liver disease become irreversible, in extreme cases, worsening to liver cancer⁴. In liver cancer (Hepatocellular Carcinoma, HCC), liver function can worsen to a state of complete liver failure, ultimately requiring transplantation⁴.

1.2 Recurrent coding mutations in liver cancer

On the genetic level, previous research has found that the most frequently occurring coding mutations in liver cancer occur in the following genes: catenin beta 1 (CTNNB1), axin 1 (AXIN1), tumor protein p53 (TP53), WW domain containing E3 ubiquitin protein ligase 1 (WWP1), and AT-rich interaction domain 1A (ARID1A)². It is hypothesized that they are

frequently mutated due to their respective role in genomic instability prevention². Specifically, genomic instability is a hallmark of many cancers, and results from mutations in DNA repair genes which ultimately drives cancer progression. Liver cancer can also be caused by long-term infection of Hepatitis B (HBV) and Hepatitis C (HCV) viruses, obesity and excessive alcohol consumption¹. Viral proteins in HBV and HCV are speculated to utilize signaling programs that cause cancer by increasing cellular proliferation and genomic instability¹.

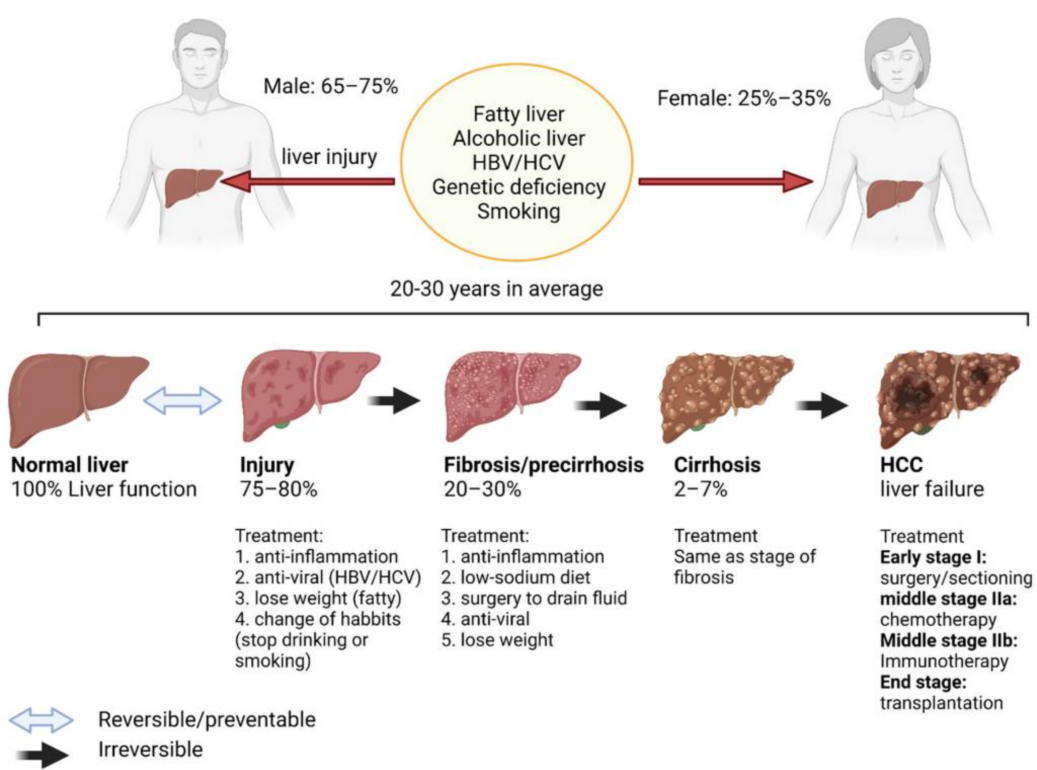


Figure 1.1 Stages of degradation of Liver function leading to Hepatocellular Carcinoma (HCC) and the respective treatment options for each stage of liver damage⁴.

1.3 Mutations in Gene regulatory elements in liver cancer

While we have a good understanding of the gene coding mutations that lead to cancer, we have a limited understanding of how cancers use this noncoding mutational landscape in order to regulate or promote cancer. Our noncoding genome space is made of regions that regulate gene

expression, which comprises about 98% of our genome. More specifically, we lack an understanding of specific mutations in gene regulatory elements, such as promoters, enhancers, and 3' untranslated regions (UTRs) that could be cancer driver mutations. Dissecting this is vital, as gene regulatory elements often mediate tissue-specific effects and can also be dysregulated in cancer⁵.

Mutations in gene regulatory elements can be a major cause of liver cancer. Some noncoding mutations in liver cancer have been associated with long noncoding RNA (lncRNA) genes, such as HOXA cluster antisense RNA 2 (HOXA-AS2) which has been shown to promote cell proliferation in liver cancer cells, as well as lncRNA SNHG17 that has been linked to large tumor size, poor differentiation, and angiogenesis in liver cancer⁶⁻⁸ with both lncRNAs encapsulating hallmarks of cancer^{9,10}. In addition, genomic structural variants that are thought to change gene expression by altering the interactions between enhancers and their target genes have been also shown to be associated with liver cancer².

Putative regulatory elements for liver cancer have been previously annotated genome-wide using technologies such as DNase-seq and histone ChIP-seq¹¹. In addition, functional mutations that cause changes in promoter activity have been identified for liver cancer. For example, activating mutations in the promoter of the Telomerase Reverse Transcriptase (TERT) were shown to be mutated in 60% of liver cancer cases². Additionally, mutations in the promoter of RAS like proto-oncogene A (RALA), which is required for the transformation of Ras mutation-driven cancer cells, are associated with tumor aggressiveness and poor patient prognosis¹². However,

other liver-cancer associated mutations in gene regulatory elements, particularly ones in enhancers, have remained largely uncharacterized.

1.4 Super enhancer landscapes differ between liver cancer and normal cells

Cancer is a complex disease of acquired somatic mutations. During tumorigenesis, DNA mutations, indels, chromosomal rearrangements, 3D chromatin structural changes and viral infections all work together to regulate the creation of oncogenic super enhancers, or clusters of enhancers in close proximity that drive transcription in the cells that acquire them¹³.

Interestingly, epigenetic studies have shown that super enhancer-targeted therapy contributes to disease prognosis¹⁴. Importantly, gene transcription dramatically decreases in the absence of enhancers¹⁵. Dysregulated activation of super enhancers has been associated with occurrence and development of liver disease, an over-arching category which includes liver cancer. Furthermore, enhancer biology is flexible because enhancer-related transcription does not depend on the direction or position of the enhancer, for example, enhancers can reside in the 5' or 3' regulatory regions, as well as in introns, and can even be located on different chromosomes. A few examples in literature have shown enhancers can undergo aberrant regulation to promote cancer by de novo enhancer assembly¹⁶. This mechanism is driven by cancer-enabling mutations to promote the cancer program forward.

Previous research has also shown that super enhancer landscapes differ between HCC and normal liver cells, specifically HCC cells are reprogrammed during the development of liver cancer¹⁷. Vital components of the super enhancer complex including bromodomain containing 4 (BRD4), cyclin-dependent kinase 7 (CDK7) and mediator complex subunit 1 (MED1) are frequently overexpressed in HCC patients and linked to poor prognosis¹⁷. In this capacity, BRD4

acts as a specific activator capable of recruiting positive transcription elongation factor b (P-TEFb), a requirement to interact with RNA polymerase II, in order to facilitate transcription¹⁸. BRD4 can also regulate DNA replication, the cell cycle, gene transcription and other cellular activities¹⁹, such that in HCC patients, de novo super enhancer formation is accompanied by BRD4 redistribution¹⁷, such that BRD4 is significantly upregulated in HCC patients and in liver cancer cell lines and its overexpression is linked to poor prognosis in HCC patients²⁰.

1.5 Utilizing neomers to identify noncoding recurrent mutations in liver cancer

In terms of 3' untranslated regions (UTRs), it was found that 3' UTR splicing is upregulated in cancer²¹. Genomic variations in 3' UTRs can disrupt regulatory elements resulting in the aberrant regulation of 3'UTR processing²². In fact, several analyses of 3' UTR-related quantitative trait loci (QTL), comprising RNA editing QTLs and 3' alternative polyadenylation (APA) QTLs have found that many 3'UTRs are dysregulated in cancer and have been associated with overall patient survival and drug response²³. These genomic variations in 3'UTRs, in addition to dysregulated 3'UTR processing events have been shown to alter the transcriptional landscape, and as such, have been linked to cancer²³.

Furthermore, recent studies have shown that somatic mutations that interrupt regulatory elements can activate oncogenes in cancer², further suggesting that gene regulatory element mutations can lead to cancer³. A major hurdle in finding these mutations includes that while promoters are nearby their gene targets, enhancers and their target gene are hard to detect and can be far away from the gene they regulate⁵ Enhancers have been shown to regulate expression of their target

gene products in a spatial and temporal manner, acting on gene targets as far as 1,000,000 basepairs (bp) away, making the functional effects of enhancers often difficult to study.



Figure 1.2 Diagram describing how neomers, DNA sequences absent from the human genome of healthy patients, can be used as a tool to dissect somatic mutations that may be associated with cancer. Somatic mutations in cancer can lead to the appearance of 16bp sequences that are generally absent from health individuals, which could be leveraged for cancer detection.

As cancer is a disease of acquired somatic mutations, we compare all possible single nucleotide substitutions in the reference genome that create the possibility of giving rise to neomers, or DNA sequences absent in healthy individuals but arising in cancer patients. While neomers range from 13-17 bp DNA sequences in length, we focus on neomers of 16bp as they represent the shortest kmer length that we were able to detect a sufficient number of recurrent neomers per patient²⁵. We filter out common germline single nucleotide polymorphisms (SNPs) and normal-recurring sequences that are not present in liver cancer patients (**Figure 1.2**). In order to distinguish between cancer types, our lab previously found that each cancer type has a distinct neomer signature, resulting in patients that are strongly enriched for a single cancer,²⁵ including liver cancer neomers that we use here. Specifically, in this manuscript, we annotate all the liver

cancer-specific neomers that reside in noncoding regions to identify cancer-enabling or driver mutations that may lead to liver cancer onset or progression (**Figure 1.2**).

While the absence of neomers could be due to chance, we and others have shown that a significant proportion of them is under negative selection pressure, suggesting that they may have a deleterious effect on the genome²⁷. In addition, Alileche et al showed that two genomic absent 5-amino acid peptides can lead to lethality in cancer cells²⁸, and later showed that the same minimal absent words have a broad lethal effect on cancer cell lines derived from nine organs²⁹. In addition, we have also shown that these sequences could be used as DNA ‘fingerprints’ to identify specific human populations²⁷. As neomers generally do not exist in a human genome, their appearance due to mutagenesis followed by clonal expansion could be exploited as a diagnostic method for diseases associated with a mutational burden, such as cancer.

Here, I characterized noncoding neomers in liver cancer that fall into promoters, enhancers, and 3’UTRs. I then use tumor-normal matched DNA sequencing information to understand which variants/mutations result in liver cancer progression. Furthermore, I use the genomic locations and map them back to gene regulatory sequences where they appear. More specifically, our lab has annotated tumor-specific neomers from over 2,500 genomes in 21 different cancer types, utilizing The Cancer Genome Atlas (TCGA) datasets²⁴ finding thousands of neomers that are specific for each cancer type²⁵. Here, I use computational tools to further characterize neomers which are more likely to be cancer driver mutations in liver cancer.

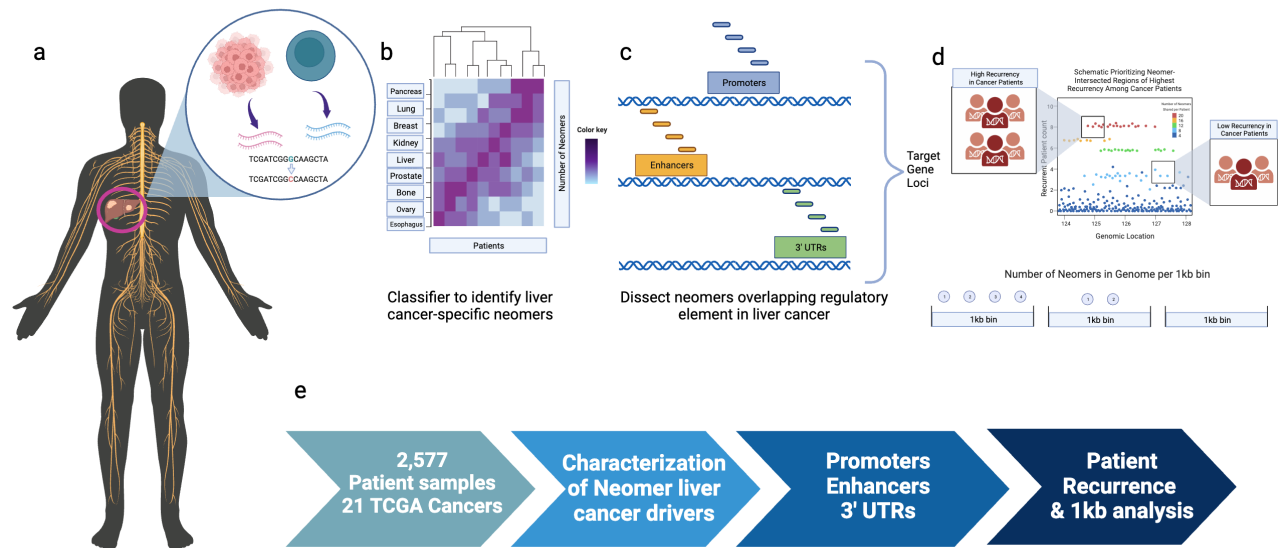


Figure 1.3 Schematic overview of neomer characterization in candidate cis-regulatory elements. (a) We identified cancer specific neomers from The Cancer Genome Atlas (TCGA) pan-cancer cohort²⁴. (b) We then used neomers to generate a cancer type classifier, which we use to identify liver cancer-specific neomers. (c) We dissect neomers that arise in gene regulatory elements such as promoters, enhancers and 3' UTRs, and assign gene targets to them using ORegAnno, a curated database for gene regulatory annotation²⁶. (d) Schematic of two forms of analysis that we utilize: (1) recurrence in patients, defined by genomic regions that have the highest patient counts, and (2) enrichment of neomers within a 1kb bin to determine, both used as supplementary metrics to dissect neomer hotspots in liver cancer patients, putatively marking regions of increased susceptibility in liver cancer patients. (e) Summary of overall characterization.

I further annotate these sequences for those more likely to be in gene regulatory elements, associated with liver cancer, using the following (**Figure 1.3**): 1) I intersect neomers with relevant functional genomics datasets that characterize gene regulatory elements (from Encode Screen V3 Registry³⁰, which is based on ChIP-seq, ATAC-seq, DNase-seq experiments) and 2) I utilized two types of analysis: enrichment of neomers within a 1 kilobase (kb) bin, and recurrence in patients in order to identify and characterize genomic hotspots of gene regulatory activity that drive cancer. More specifically, neomer hotspots showing a higher recurrence in liver cancer patients may suggest that they are likely to be associated with cancer.

1.6 References

1. Choi, S. et al. Global Burden of Primary Liver Cancer and its Association with Underlying Aetiologies, Sociodemographic Status, and Sex Differences from 1990-2019: A DALY-based Analysis of the Global Burden of Disease 2019 Study. *Clin. Mol. Hepatol.* **29**, 433–452 (2023).
2. Rao, C. V., Asch, A. S. & Yamada, H. Y. Frequently mutated genes/pathways and genomic instability as prevention targets in liver cancer. *Carcinogenesis* **38**, 2–11 (2017).
3. “Liver Cancer Symptoms.” MD Anderson Cancer Center, www.mdanderson.org/cancer-types/liver-cancer/liver-cancer-symptoms.html.(2023).
4. Wu, Bocheng, et al. “Inflammation, Fibrosis and Cancer: Mechanisms, Therapeutic Options and Challenges.” *Cancers (Basel)* **14**(3): 552 (2022).
5. Herz, H.-M., et al. Enhancer Malfunction in Cancer. *Mol Cell*, **53**, 859-866 (2014).
6. Khan, A. & Zhang, X. Function of the Long Noncoding RNAs in Hepatocellular Carcinoma: Classification, Molecular Mechanisms, and Significant Therapeutic Potentials. *Bioengineering (Basel)* **9**, (2022).
7. Kim, Y. -A., Park, K.-K., and Lee, S.-J. LncRNAs Act as a Link between Chronic Liver Disease and Hepatocellular Carcinoma. *Int. J. Mol. Sci.* **21**, (2020).
8. Luo, Y. et al, LncRNA SNHG17 Contributes to Proliferation, Migration, and Poor Prognosis of Hepatocellular Carcinoma. *Can J Gastroenterol Hepatol* **2021**, 9990338 (2021).

9. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: the Next Generation. *Cell* **144**, 646-674 (2011).
10. Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov.* **12**, 31-46 (2022).
11. Lowdon, R. F. & Wang, T. Epigenomic Annotation of Noncoding Mutations Identifies Mutated Pathways in Primary Liver Cancer. *PLoS One* **12**, e0174032 (2017).
12. Tian, L. et al. Dysregulation of RalA Signaling through Dual Regulatory Mechanisms Exerts its Oncogenic Functions in Hepatocellular Carcinoma. *Hepatology* **76**, 48-65 (2022).
13. Jia Q, Chen S, Tan Y, Li Y, Tang F. Oncogenic Super-Enhancer Formation in Tumorigenesis and its Molecular Mechanisms. *Exp Mol Med.* **52**(5):713-723 (2020).
14. Shin HY. Targeting Super-Enhancers for Disease Treatment and Diagnosis. *Mol Cells.* **41**(6):506-514 (2018).
15. Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G. Enhancers: Five Essential Questions. *Nat Rev Genet.* **14**(4):288-95 (2013).
16. Mansour MR, Abraham BJ, Anders L, Berezovskaya A, Gutierrez A, Durbin AD, Etchin J, Lawton L, Sallan SE, Silverman LB, et al. An Oncogenic Super-Enhancer Formed through Somatic Mutation of a Noncoding Intergenic Element. *Science.* **346**:1373–1377 (2014).
17. Tsang FH, Law CT, Tang TC, Cheng CL, Chin DW, Tam WV, Wei L, Wong CC, Ng IO, Wong CM. Aberrant Super-Enhancer Landscape in Human Hepatocellular Carcinoma. *Hepatology.* **69**(6):2502-2517 (2019).

18. Jang MK, Mochizuki K, Zhou M, Jeong HS, Brady JN, Ozato K. The Bromodomain Protein Brd4 is a Positive Regulatory Component of P-TEFb and Stimulates RNA Polymerase II-Dependent Transcription. *Mol Cell*. **19**(4):523-34 (2005).
19. Tang X, Peng R, Ren Y, Apparsundaram S, Deguzman J, Bauer CM, Hoffman AF, Hamilton S, Liang Z, Zeng H, Fuentes ME, Demartino JA, Kitson C, Stevenson CS, Budd DC. BET Bromodomain Proteins Mediate Downstream Signaling Events Following Growth Factor Stimulation in Human Lung Fibroblasts and Are Involved in Bleomycin-Induced Pulmonary Fibrosis. *Mol Pharmacol*. **83**(1):283-93 (2013).
20. Zhang P, Dong Z, Cai J, Zhang C, Shen Z, Ke A, Gao D, Fan J, Shi G. BRD4 Promotes Tumor Growth and Epithelial-Mesenchymal Transition in Hepatocellular Carcinoma. *Int J Immunopathol Pharmacol*. **28**(1):36-44 (2015).
21. Chan, Jia Jia, et al. Pan-Cancer Pervasive Upregulation of 3' UTR Splicing Drives Tumourigenesis. *Nature Cell Biology*, **24**, 928–939 (2022).
22. Hong D, Jeong S. 3'UTR Diversity: Expanding Repertoire of RNA Alterations in Human mRNAs. *Mol Cells*. **46**(1):48-56 (2023).
23. Chan, Jia Jia, et al. 3'UTR Heterogeneity and Cancer Progression. *Trends in Cell Biology*, **33**(7):568-582 (2023).
24. ICGC/TCGA Pan Cancer Analysis of Whole Genomes Consortium. Pan-Cancer Analysis of Whole Genomes. *Nature* **578**, 82-93 (2020).

25. Georgakopoulos-Soares, I. et al, Leveraging Sequences Missing from the Human Genome to Diagnose Cancer. medRxiv (2023) doi: 10.1101/2021.08.15.21261805
26. Lesurf, R. et al. ORegAnno 3.0: A Community-driven Resource for Curated Regulated Annotation. *Nucleic Acids Res.* **44**, D126-32 (2016).
27. Georgakopoulos-Soares, Ilias, et al. “Absent from DNA and Protein: Genomic Characterization of Nullomers and Nullpeptides across Functional Categories and Evolution - Genome Biology.” *BioMed Central.* (2021).
28. Alileche A., Goswami J., Bourland W., Davis M., Hampikian G.. Nullomer Derived Anticancer Peptides (NulloPs): Differential Lethal Effects on Normal and Cancer Cells in vitro. *Peptides.* **38**:302–311 (2012).
29. Alileche A., Hampikian G.. The Effect of Nullomer-derived Peptides 9R, 9S1R and 124R on the NCI-60 Panel and Normal Cell Lines. *BMC Cancer.* **17**:533 (2017).
30. ENCODE Project Consortium et. al. Expanded Encyclopaedias of DNA Elements in the Human and Mouse Genomes. *Nature* **583**, 699-710 (2020).

Chapter 2: Utilizing short sequences missing from the genome to identify gene regulatory element mutations associated with liver cancer

Jasmine Sims^{1,2}, Ofer Yizhar Barnea^{1,2}, Chengyu Deng^{1,2}, Weiyu Li^{1,2}, Jingjing Zhao^{1,2}, Ryder Easterlin^{1,2}, Ilias Georgakopoulos-Soares^{1,2,3}, Franklin Huang^{2,4,5,6,7,8,9}, Martin Hemberg¹⁰, Nadav Ahituv^{1,2^}

¹Department of Bioengineering and Therapeutic Sciences, University of California San Francisco, San Francisco, California, USA.

²Institute for Human Genetics, University of California San Francisco, San Francisco, California, USA.

³Institute for Personalized Medicine, Department of Biochemistry and Molecular Biology, The Pennsylvania State University College of Medicine, Hershey, PA, USA.

⁴Bakar Computational Health Sciences Institute

⁵C3 (Computational Cancer Community) Initiative

⁶UCSF Department of Medicine

⁷Helen Diller Family Comprehensive Cancer Center

⁸San Francisco Veterans Affairs Medical Center

⁹UCSF Department of Medicine

¹⁰Evergrande Center for Immunologic Diseases, Harvard Medical School and Brigham and Women's Hospital, Boston, USA.

[^] Corresponding author: nadav.ahituv@ucsf.edu

2.1 Summary

Liver cancer is diagnosed yearly in over 800,000 people worldwide. While several gene coding mutations have been found to be associated with this cancer, we lack an understanding of gene regulatory driver mutations that could also lead to liver cancer. Using short sequences, 16 base pairs in length, termed neomers, that are generally absent from the human genome but appear in liver cancer, we identify numerous potential causative gene regulatory mutations in promoters, potential enhancers and 3'UTRs. We identify noncoding neomers that are enriched in numerous patients and are thought to regulate cancer-associated genes. We also characterize specific one kilobase regions in the human genome that are enriched for noncoding neomers that reside near many relevant liver cancer genes. Combined, our study identifies novel mutations in gene regulatory elements that could be key regulators of liver cancer, providing a basis for potential downstream diagnosis and therapeutics.

2.2 Introduction

Liver cancer is the seventh most common form of cancer¹, with over 800,000 people worldwide estimated to be diagnosed for liver cancer each year, and 700,000 passing away from liver cancer per year¹. Genetic factors associated with liver cancer include metabolic syndrome, diabetes, and male gender, with males being more prone than females². On the genetic level, previous research has found that the most frequently occurring coding mutations in liver cancer occur in the following genes: catenin beta 1 (CTNNB1), axin 1 (AXIN1), tumor protein p53 (TP53), WW domain containing E3 ubiquitin protein ligase 1 (WWP1), and AT-rich interaction domain 1A (ARID1A)². It is hypothesized that they are frequently mutated due to their respective role in genomic instability prevention². Liver cancer can also be caused by long-term infection of Hepatitis B (HBV) and Hepatitis C (HCV) viruses, obesity and excessive alcohol consumption¹. Viral proteins in HBV and HCV are speculated to utilize signaling programs that cause cancer by increasing cellular proliferation and genomic instability¹. In terms of liver cancer associated noncoding mutations, our current understanding is limited. Some mutations in liver cancer have been associated with long noncoding RNA (lncRNA) genes, such as HOXA cluster antisense RNA 2 (HOXA-AS2) which has been shown to promote cell proliferation in liver cancer cells, as well as lncRNA SNHG17 that has been linked to large tumor size, poor differentiation, and angiogenesis in liver cancer³⁻⁵ with both lncRNAs encapsulating hallmarks of cancer^{6,7}. In addition, genomic structural variants that are thought to change gene expression by altering the interactions between enhancers and their target genes have also been associated with liver cancer². However, we currently lack an understanding of specific mutations in gene regulatory elements that could be cancer driver mutations.

Mutations in gene regulatory elements can also be a major cause of liver cancer. Gene regulatory elements, such as promoters, enhancers, silencers, and insulators, often mediate tissue-specific effects and can also be dysregulated in cancer⁸. Putative regulatory elements for liver cancer have been previously annotated genome-wide using technologies such as DNase-seq and histone ChIP-seq⁹. In addition, functional mutations that cause changes in promoter activity have been identified for liver cancer. For example, activating mutations in the promoter of the Telomerase Reverse Transcriptase (TERT) were shown to be mutated in 60% of liver cancer cases². Additionally, mutations in the promoter of RAS like proto-oncogene A (RALA), which is required for the transformation of Ras mutation–driven cancer cells, are associated with tumor aggressiveness and poor patient prognosis¹⁰. However, other liver-cancer associated mutations in gene regulatory elements, particularly ones in enhancers, have remained largely uncharacterized.

Neomers are short DNA sequences, 13-17 base pairs (bp) in length, that are generally missing from the human genome of healthy individuals, but present in individuals with cancer¹¹. In a previous study, we have shown that these sequences can be used to diagnose cancer from cell free DNA (cfDNA) and identified potential cancer regulatory element driver mutations in prostate cancer¹¹. For the latter, we used massively parallel reporter assays (MPRA) that can test thousands of sequences/variants for their regulatory activity by placing them in front of a minimal promoter and transcribed barcode¹². Here, we set out to identify gene regulatory element driver mutations associated with liver cancer. We used neomers that we previously annotated for liver cancer¹¹ and characterized them. We first characterize noncoding neomers associated with liver cancer, including promoters, enhancers and 3' untranslated regions (3'UTRs) finding several to regulate known liver cancer associated genes. We next analyzed

neomers for patient recurrence finding that our top recurring neomers are thought to drive expression in tumor suppressors and signaling pathway intermediates. Furthermore, we binned the human genome to 1 kilobase (kb) regions, to annotate genomic regions with high neomer density, identifying several high neomer density 1kb regions near known liver associated genes, such as the gene Shroom family member 2 (SHROOM2) and the small nucleolar RNA H/ACA (SNORA31), which are significantly associated with tumor size and stage¹³. Combined, our work identified novel gene regulatory mutations that could be associated with liver cancer that can be utilized for diagnosis and therapy.

2.3 Results

We previously annotated 4,174,421 liver cancer associated 16bp neomers from The Cancer Genome Atlas (TCGA) database^{11,14}. These unique 16bp neomers were found only in the 336 liver cancer patients and not in their matching control tissue. We used neomers 16bp in length, as this length allowed us to detect a sufficient number of neomers per patient¹¹. Annotation of these neomers found 19,815 neomer mutations (0.5%) to reside in protein coding sequences and 4,154,606 (99.5%) liver cancer associated neomers in noncoding regions. We next set out to characterize these noncoding liver cancer associated neomers.

2.3.1 Annotation of liver cancer-associated neomers in promoters

As the target gene of the promoter can be easily identified due to their genomic proximity to the gene's transcription start site (TSS), we first analyzed the genes whose promoters encompassed liver cancer associated neomers. We found liver cancer-associated neomers in the promoters of 4,056 genes, as defined by ENCODE SCREEN¹⁵, that uses methylation, ChIP-seq for transcription factors and histone marks and open chromatin to annotate promoters in human hepatocellular carcinoma cells (HepG2). We carried out GO term enrichment for these genes using clusterProfiler¹⁶, finding enrichment for 'regulation of chromosomal organization', 'RNA splicing', mitosis and various muscle developmental GO terms (**Figure 2.1A; Supplementary Table 1**).

Genes involved in the 'regulation of chromosomal organization' include the MYC proto-oncogene, bHLH transcription factor (MYC), ATRX chromatin remodeler, alpha thalassemia/mental retardation syndrome X-linked (ATRX), cell division cycle associated 5

(CDCA5) and cell division cycle 6 (CDC6). MYC undergoes significant changes in tumor proliferation, metastasis, and plays a major role in resistance to comprehensive therapies¹⁷. ATRX is a frequently mutated tumor suppressor gene in liver cancer and acts as a chromatin remodeler and a transcriptional regulator on the protein level, such that disruption of ATRX is thought to contribute to the development of cancer and resistance to cancer treatment¹⁸. CDCA5 promotes oncogenesis and has a vital role in the interaction between cohesin and chromatin in interphase, and its aberrant expression in many cancers, including liver cancer, acts as a biomarker for poor prognosis¹⁹. CDC6 also acts as a biomarker in liver cancer, and its expression is upregulated in liver cancer compared to normal liver tissue²⁰.

Genes associated with 'RNA splicing' include RNA, U5E small nuclear 1 (RNU5E-1), RNA, U6 small nuclear 7 (RNU6-7), pre-mRNA processing factor 19 (PRPF19). Lower expression of RNU5E-1 is independently linked to severe tumor-free and long-term survival in liver cancer²¹, while PRPF19 has been shown to aid in colorectal cancer metastasis to the liver via the Src-YAP1 pathway²². Genes associated with mitotic metaphase/anaphase transition included RB transcriptional corepressor 1 (RB1), aurora kinase B (AURKB), APC regulator of WNT signaling pathway (APC). RB1 acts as a tumor suppressor and functions by coordinating multiple pathways that impact cancer initiation and progression²³. AURKB is a major mitotic protein kinase important for cell cycle progression and is used as a prognostic biomarker in liver cancer²⁴. Loss of APC is known to act synergistically with oncogenes to induce liver tumor formation²⁵.

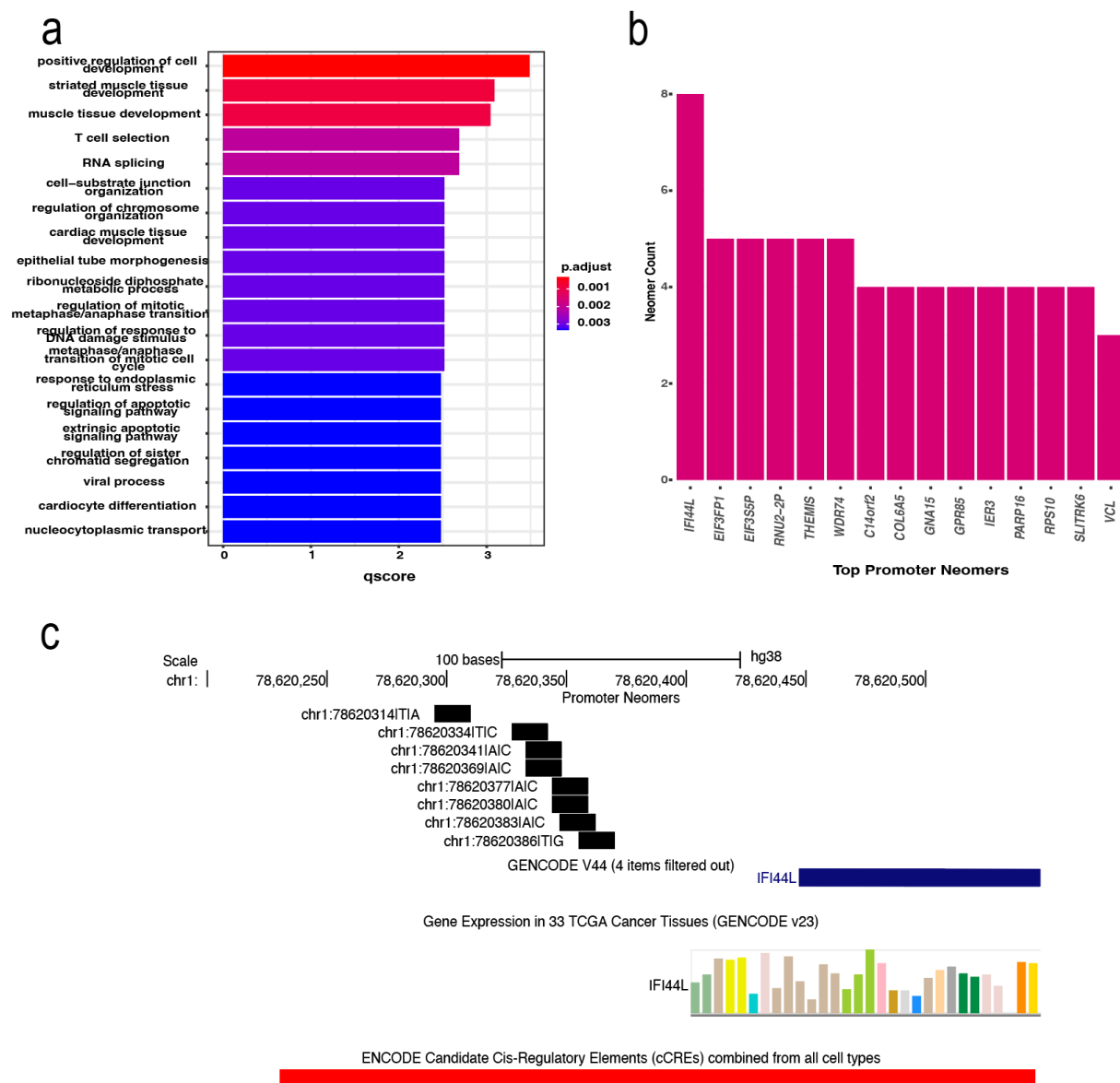


Figure 2.1 Characterization of promoter neomers in liver cancer patients. **(a)** Comparison of gene ontology (GO) terms specific to neomers that overlap ENCODE SCREEN defined promoter elements. p-value cutoff=0.05, adjusted p-value defined by Benjamini-Hochberg. QScore values were calculated by the $-\log$ base 10 of the adjusted p-value. **(b)** Bar chart of the gene promoters that have the most neomer mutations across liver cancer patients. **(c)** UCSC genome browser snapshot of the IFI44L promoter with genomic tracks showing neomers, GENCODE V44 track, TCGA cancer gene expression with the brown color depicting liver cancer and ENCODE candidate cis-regulatory elements (cCREs).

We next ranked the promoter neomers by the number of neomers that arise in different sequence coordinates in that promoter (**Figure 2.1B**; **Supplementary Table 2**). Our top promoter, having

8 neomers, regulates the Interferon Induced Protein 44 Like (IFI44L) gene (**Figure 2.1C**). IFI44L is a tumor suppressor that has been shown to affect cancer stemness, metastasis and drug resistance via activation of the MET/Src signaling pathway²⁶. We found 5 neomers in the promoter of the thymocyte selection associated (Themis) gene. Themis acts as a T-cell lineage-specific protein that is responsible for regulating T-cell selection in the thymus by setting the threshold for positive and negative selection²⁷, with Themis^{-/-} CD4⁺ T cells resulting in reduced proliferation compared to Themis^{+/+} CD4⁺ T cells²⁷. Additionally, we found 4 neomers in the promoter of Immediate Early Response 3 (IER3) gene. IER3 was found to have significantly increased expression in liver cancer, detected by immunohistochemistry (IHC)²⁸, and has been associated with tumor progression in liver cancer²⁸.

2.3.2 Annotation of liver cancer associated neomers in enhancers

We next analyzed how many of these noncoding neomers overlap potential enhancers as defined by ENCODE SCREEN¹⁵. We found neomers to overlap 33,265 distal enhancers and 5,969 proximal enhancers. We next linked distal and proximal enhancers to their target genes using OregAnno³⁰, a tool that links candidate regulatory elements to their target genes. We then utilized these target genes to perform GO analysis using clusterProfiler³¹, finding an enrichment for ‘axon development’ and ‘mesenchymal cell differentiation’ (**Figure 2.2A; Supplementary Table 2**), which is associated with a cellular state transition in cancer cells in metastatic cancers³². For the axon developmental GO terms, liver hepatocellular carcinoma (LIHC) can be divided into two subtypes, based on their expression of neural-related genes, with prognosis and stage differences between the subtypes that express neural-related genes, and the subtype that lacks neural-related gene expression³³, likely leading to this enrichment. Genes enriched in the mesenchymal cell differentiation GO term include: transforming growth factor beta 2 (TGFB2),

twist family bHLH transcription factor 1 (TWIST1) and notch receptor 1 (NOTCH1)(**Figure 2.2A**). TGFB2 has been shown to be upregulated in hepatocellular carcinoma (HCC) and correlated with liver cancer development in mice³⁴. TWIST1 is a prognostic biomarker of liver cancer, associated with invasion, migration and epithelial-mesenchymal transformation (EMT) of HCC cells⁷. NOTCH1 regulates the c-Jun N-terminal kinase (JNK) signaling pathway and increases apoptosis in HCC⁸.

We also found an enrichment for the regulation of small GTPase mediated signal transduction, representing 204 out of 7879 genes (p-value 3.261387×10^{-19})(**Figure 2.2A; Supplementary Table 2**). Small GTPases, such as RHO GTPases, have been implicated in cancer progression including proliferation, migration, invasion and metastasis of liver cancer³⁵. Genes enriched in the small GTPase included RAS p21 protein activator 3 (RASA3), SOS Ras/Rac guanine nucleotide exchange factor 1 (SOS1) and mitogen-activated protein kinase kinase kinase 4 (MAP4K4)(**Figure 2.2B**). Oncogenic forms of Ras, such as RASA3, are involved in ~30% of all human cancers, including liver cancer³⁶. Additionally, RASA3 is frequently hypomethylated in liver cancer, and has been suggested to be a putative biomarker in early detection of liver cancer³⁷. SOS1 is involved in invasion and metastasis of liver cancer cells, and is overexpressed in liver cancer patients, and correlates with poor prognosis in liver cancer³⁸. MAP4K4 is overexpressed in many types of cancer, including liver cancer, and has been shown to be an independent predictor of poor prognosis in liver cancer patients³⁹. Additionally, we found enrichment for the Wnt signaling pathway, representing 282 out of 7879 genes (p-value 1.442951×10^{-18})(**Figure 2.2A; Supplementary Table 2**). WNT signaling has been implicated in the malignant transformation and expansion of liver cells, leading to metastatic liver cancer⁴⁰. Enriched genes in the WNT pathway include: Wnt family member 7A (WNT7A), Wnt family

member 11 (WNT11), and SMAD family member 3 (SMAD3)(**Figure 2.2A**). Overexpression of WNT7A has been shown to inhibit tumor cell migration⁴¹. Overexpression of WNT11 is thought to activate Rho kinase, resulting in the reduction of liver cancer cell migration⁴². SMAD3 acts as a tumor suppressor by causing apoptosis of TGF- β in liver cancer⁴².

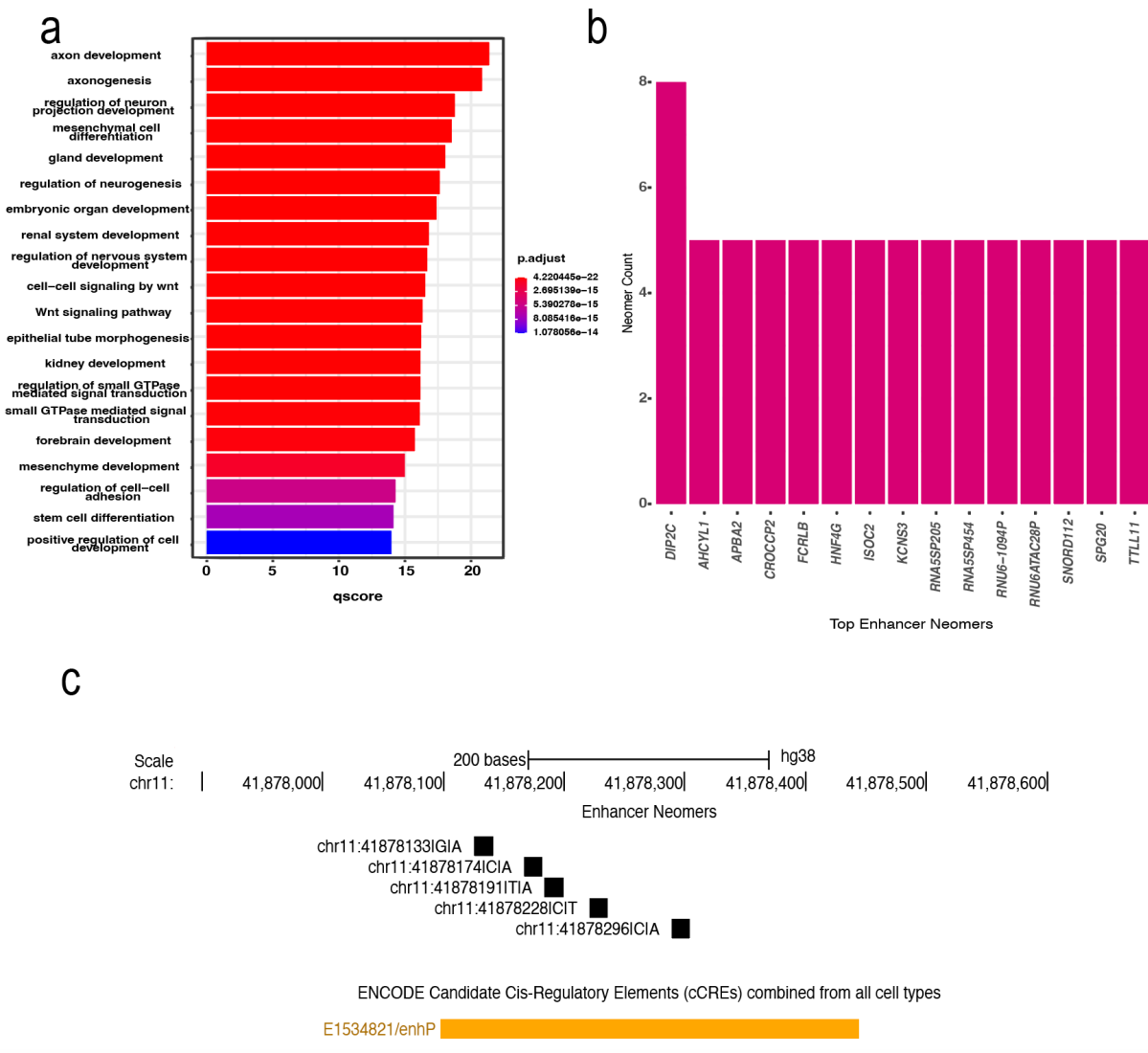


Figure 2.2 Liver cancer-associated neomers in potential enhancers. **(a)** Comparison of gene ontology (GO) terms specific to neomers that overlap ENCODE SCREEN defined enhancer elements. **(b)** Bar chart of top enhancer target gene loci that carry the highest neomer density across liver cancer patients. **(c)** UCSC genome browser snapshot of enhancer neomer tracks overlapping ENCODE candidate cis-regulatory elements (cCREs).

We next searched for noncoding ENCODE SCREEN¹⁵ enhancer elements that harbor the most neomer mutations that arise in different sequence coordinates in that SCREEN element (**Supplementary Table 2**). Our top candidate enhancer has 8 neomers, and is suggested by OregAnno³⁰ to regulate the disco interacting protein 2 homolog C (DIP2C) gene. While DIP2C has yet to be associated with liver cancer, previous work has shown that tumor-derived extracellular vesicle miR-375 directly targets DIP2C, which upregulates the WNT signaling pathway, resulting in prostate cancer bone metastasis⁴³. Additionally, we find a potential enhancer sequence with 5 neomers, which is thought to regulate the Fc receptor like B (FCRLB) gene (**Supplementary Table 2**). High expression of FCRL genes has been linked with a protective effect in many cancers⁴⁴. FCRLB has been shown to be a risk factor in many types of cancer, including brain and lung cancer, but has yet to be associated with liver cancer as a risk factor⁴⁴. Notably, FCRLB has been shown to have significantly lower expression in liver cancer compared to normal tissue and positively correlated with copy number variation (CNV) in several tumors⁴⁴. Furthermore, we found a candidate enhancer with 5 neomers that is suggested to regulate the hepatocyte nuclear factor 4 gamma (HNF4G) gene. Knockdown of HNF4G inhibits cancer cell proliferation and has been correlated with poor prognosis, resulting in tumor cell growth in colorectal cancer, and has been shown to facilitate proliferation and invasion of bladder cancer⁴⁵, but has yet to be associated with liver cancer.

2.3.3 Liver cancer associated neomers in the 3' untranslated region

We next analyzed neomers that fall within the 3' untranslated region (UTR), which is important for mRNA stability and translation. We found 3,217 neomers in 3'UTRs targeting 1,757 genes. We carried out GO term enrichment for these genes using clusterProfiler³¹ finding enrichment for 'pattern speciation process', 'regionalization', 'stem cell differentiation' and other developmental

associated GO terms (**Figure 2.3A; Supplementary Table 3**). Neomer genes enriched in stem cell differentiation include: SRY-box transcription factor 6 (SOX6), SRY-box transcription factor 11(SOX11), and TEA domain transcription factor 2 (TEAD2). SOX6 has been shown to inhibit growth of liver cancer, where a decrease in SOX6 expression is linked to poor prognosis in liver cancer patients⁴⁶. Aberrant expression of SOX11 is associated with increased cancer cell survival and tumor metastasis and angiogenesis⁴⁷. Upregulation of expression of TEAD2 has been linked with poor prognosis in patients with liver cancer, and as such as been identified as a prognostic factor in liver cancer⁴⁸. Neomer genes enriched in mesenchyme cell development include: Phosphatase and Tensin Homolog (PTEN), Transforming Growth Factor Beta Receptor 2 (TGFBR2), and Notch Receptor 4 (NOTCH4). Many of these genes have been associated with liver cancer, in particular PTEN and TGFBR2. PTEN deficiency promotes hepatic tumor formation⁴⁹, and TGFBR2 acts as a tumor suppressor resulting in liver cancer⁵⁰.

We next examined the 3'UTRs that had the most neomers (**Figure 2.3B; Supplementary Table 4**). We found the top 3'UTR targeted to be the ALF transcription elongation factor 4 (AFF4) gene having 12 neomers in different genomic coordinates of the 3'UTR (**Figure 2.3C**). AFF4 has been shown to be associated with colorectal cancer⁵¹ but has yet to be linked to liver cancer. We also found the 3'UTR of Ras association domain family member 9 (RASSF9), to have 7 neomers. RASSF9 promotes Non-Small Cell Lung Cancer (NSCLC) cell proliferation via activation of MEK/ERK pathway⁵². While RASSF9 has yet to be linked to liver cancer, other Ras association domain family members have been shown to act as tumor suppressors in liver cancer, such as RASSF10 and RASSF5C⁵³.

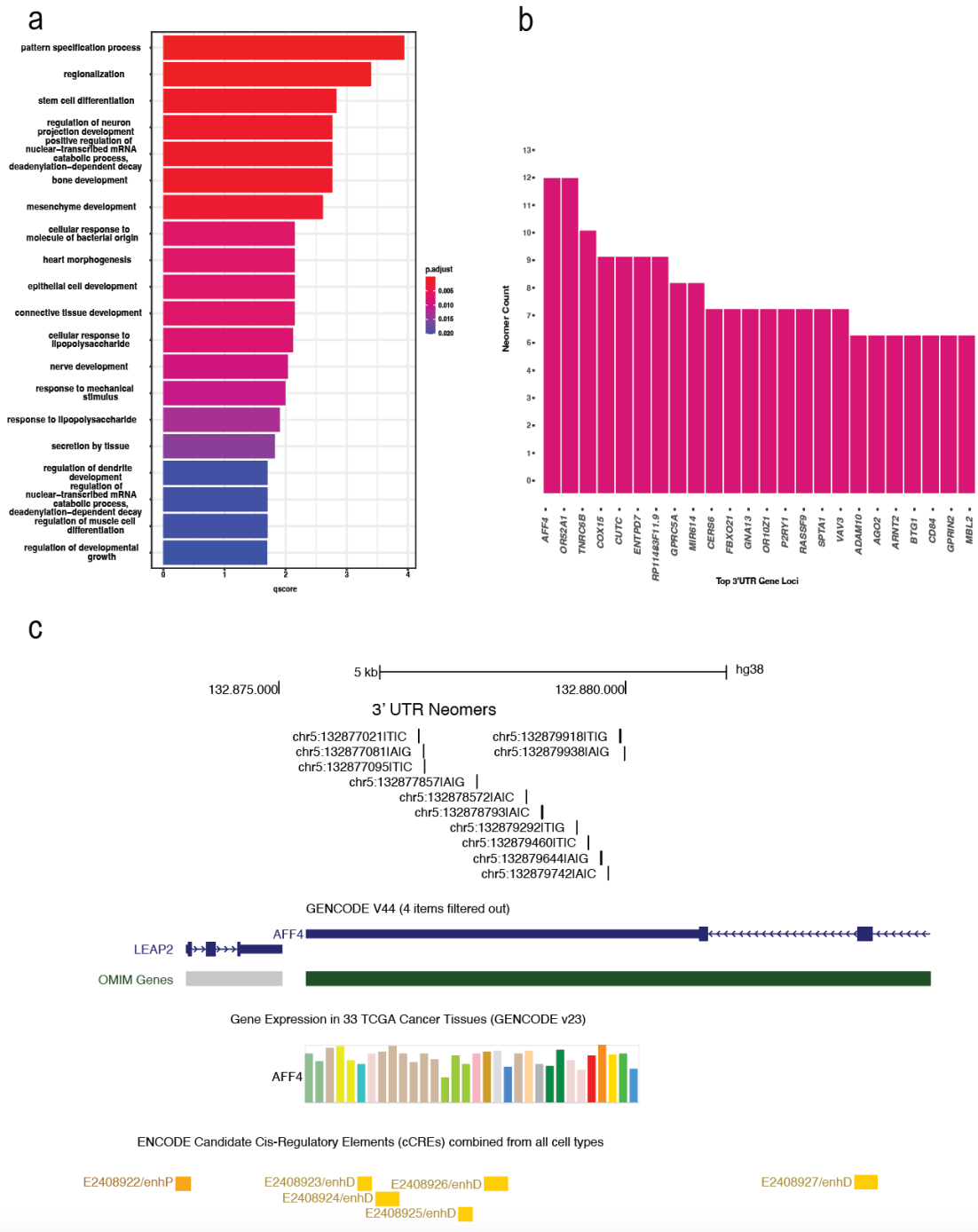


Figure 2.3 Characterization of liver cancer-associated neomers in 3' UTRs.

(a) Top 20 gene ontology (GO) terms specific to neomers that overlap hg38 knownGene 3' UTRs by transcript. (b) Bar chart of top 3' UTR genes that carry the highest neomer density across liver cancer patients. (c) UCSC genome browser snapshot of 3' UTR neomers in the AFF4 locus showing gene tracks from RefSeq, TCGA cancer gene expression with the brown color depicting liver cancer and ENCODE candidate cis-Regulatory elements (cCREs).

2.3.4 Neomers recurrent in liver cancer patients

We next set out to annotate liver cancer neomers that appear in a larger number of patients, suggesting that they could be more likely to drive cancer progression. To increase our ability to find neomers that could be associated with liver cancer, we only analyzed ones that appear in three or more patients. We also removed neomers that could appear multiple times in the same genomic position due to different mutations. We found 4,764 neomers that reside in at least three patients (**Table 2.1; Supplementary Table 4**). Our top neomer was observed in 19 patients in an intergenic region where the closest gene is the kelch like family member 13 (KLHL13), a gene which is downregulated in metastatic liver cancer⁵⁴. The second top noncoding region to have mutations in several patients (15 patients) is in an intron of the zinc finger protein 93 (ZNF93). ZNF93 is known to be involved in promoting proliferation and migration in ovarian cancer⁵⁵ but has yet to be associated with liver cancer. We also found a neomer in 12 patients in the promoter of telomerase reverse transcriptase (TERT) gene. Sixty percent of liver cancer patients are known to carry TERT promoter mutations, which lead to the expression of TERT, and subsequent telomerase elongation and cancer progression⁵⁶. Additionally, we found recurrent neomers in a potential enhancer that based on OregAnno³⁰, is thought to regulate fibroblast growth factor 12 (FGF12), with 9 patients. FGF12 has been shown to promote liver fibrosis progression⁵⁷, but has yet to be associated with liver cancer. In summary, we find that our top recurrent neomers identify candidate regulatory elements that could be associated with liver cancer.

Table 2.1 The top 10 recurrent neomers in liver cancer patients.

Neomer sequence	Coordinate(hg38)	Change	Patient count	Target loci
AGACAAGAAAATGCTG	chrX:117445366	C>A	19	NA
TGTCATCAATAATCGA	chr19:19910432	A>C	15	ZNF93
GGAGGGCCCGGAAGGG	chr5:1295113	G>A	12	TERT
AGCTCCTATCTCAAGG	chr19:20433750	G>A	10	NA
CCGATGGATCTTGACT	chr12:86038672	G>A	9	MGAT4C
TCTATATGTCTGTTTT	chr11:55394626	C>G	8	NA
ACTCTCCAATCTACCC	chrX:48164849	C>T	8	SSXP3
CTTGGTGAACTGACA	chr18:28227680	G>A	7	NA
GCCACCTCGTGGACAG	chr3:192444701	T>C	7	FGF12
ACCTACAGAATGGGAG	chr7:25450648	A>G	7	NA

2.3.5 Neomers recurrent in 1kb regions

We next set out to annotate liver cancer neomers for those that have high neomer density in 1kb regions, under the assumption that 1kb regions where there is a clustering of neomers could be more strongly associated with liver cancer. We annotated the human genome for 1kb windows (see Methods) and intersected them with our liver cancer recurrent neomers. We then ranked these 1kb bins based on bins that have the most neomers and also the most recurrence in patients (Table 2.2; Supplementary Table 5). The top 1kb region, with 79 neomers in 83 patients, is an intergenic region with the nearest gene regulatory element being a distal enhancer (EH38E2396206) and the nearest gene is ENSG00000252337, which overlaps small nucleolar RNA, H/ACA box 31 (SNORA31). SNORA31 has been shown to be downregulated in liver cancer with significant association with tumor size, capsular invasion, degree of tumor and tumor lymph node metastasis (TNM) stage¹³. In addition, lower expression of SNORA31 was associated with shorter disease-free survival and shorter overall survival in liver cancer¹³. The

next top 1kb region, with 53 neomers, affecting 59 liver cancer patients overlaps the Shroom family member 2 (SHROOM2) gene, which has been shown to inhibit tumor metastasis through both RhoA-ROCK pathway-dependent and -independent mechanisms in nasopharyngeal carcinoma⁵⁹. While SHROOM2 has yet to be linked to liver cancer, RhoA has been shown to promote liver cancer cell migration and invasion via ROCK1/MLC signaling⁶⁰. Notably, RhoA inactivation results in increased tumor growth and metastasis in liver cancer⁶¹.

Table 2.2 The top 10 1kb genomic neomer containing regions.

Chr	Start	End	Neomer count	Patient count	Target loci
chr5	106388000	106389000	79	83	NA
chrX	9792000	9793000	53	59	<i>SHROOM2</i>
chr4	31817000	31818000	51	54	NA
chr4	31815000	31816000	37	37	NA
chr3	111780000	111781000	34	37	<i>PLCXD2</i>
chr2	125894000	125895000	29	29	NA
chr19	15572000	15573000	27	27	NA
chr2	80931000	80932000	26	28	NA
chr4	31818000	31819000	26	26	NA
chr4	31816000	31817000	24	25	NA

Discussion

Mutations in gene regulatory elements can be important cancer drivers but are difficult to identify. Here, utilizing neomers, we annotated liver cancer candidate regulatory element sequence variants. We show how short DNA sequences missing from the general population can be used to identify gene regulatory elements associated with liver cancer. Utilizing these sequences, we manage to identify candidate promoters, enhancers and 3'UTRs that could be associated with liver cancer. Many of these reside near cancer-associated genes, suggesting that we find bona fide targets. However, additional functional work to characterize the effect of these mutations will be needed such as a luciferase assay or a massively parallel reporter assay to elucidate gene expression changes associated with neomers in regulatory elements.

We find that top neomers arise in regions associated with Wnt and RhoA GTPase signaling. Cancer cells experience aberrant signaling cascade programs that often lead to an alteration in the cell cycle division signal to achieve proliferative cell growth. Normally, Wnt signaling regulates self-renewal of stem cells and the ability to differentiate, but when aberrantly regulated, promotes the progression of cancer stem cells, leading to cancer metastasis⁶². Additionally, RhoA has been shown to control cell type-specific signaling mechanisms, which ultimately regulate cytokinesis and adhesion of Mouse Embryonic Fibroblast (MEF) cells⁶³. Notably, Rho GTPases, including RhoA, have been shown to affect cytoskeleton dynamics resulting in modulation of tumor cell aggressivity⁶⁴. In liver cancer, RHOA is a transcriptional target of the Transforming growth factor beta superfamily (TGF- β) pathway⁶⁴. Depletion of RhoA augments K-Ras driven tumor formation, functioning as a molecular brake on liver cancers that are driven by Ras, as shown in a zebrafish liver cancer model⁶⁵. In humans, RhoA is commonly upregulated in liver

cancer tissues at both the mRNA and protein levels and is associated with poor patient prognosis in liver cancer⁶⁶. Thus, neomers that are associated with the regulation of WNT and RhoA signaling pathways could be of interest for follow up functional analyses.

While we recognize that regions recurrent in liver cancer patients tend to represent an abundance of regulatory elements that regulate many known cancer-associated genes, our analysis depends on previously identified genomic loci generated by Refseq⁶⁷, The Cancer Genome Atlas (TCGA)¹³, and UCSC Genome Browser⁶⁸, and Screen V3 Registry from Encode¹⁵ to confirm sequence and genomic loci annotation identity. Due to this, this analysis favors neomers in regulatory regions that have been previously annotated and does not discover mutations that lead to the ‘birth’ of a new regulatory element. Further functional work for these sequences could potentially identify novel regulatory elements that are generated by neomer mutations.

We tiled the genome into 1kb bins in order to identify genomic regions enriched for neomers. We recognize that some regions of the genome are more hypermutable than other regions by nature, such as regions associated with somatic hypermutation associated with VDJ end-joining in T cell and B cell Receptor repertoire. Keeping these biases in mind, we then asked which 1kb bins have the most patients associated with them. By reanalyzing the data in this way, we could find genomic hotspots most targeted in liver cancer patients. We also found several of these regions to cluster in specific genomic regions, highlighting these regions to be associated with liver cancer. We propose that neomers could influence gene regulation in liver cancer by arising as mutations in gene regulatory elements, resulting in aberrant gene regulation patterns that contribute to cancer progression. Further cancer whole-genome sequencing datasets, gene regulatory element

functional genomic characterization and regulatory element functional characterization will allow to characterize the effect of these neomers on cancer in the future.

Methods

Genomic Feature Annotation

We characterized coding neomers as those that overlap the consensus coding (CDS) using RefSeq⁶⁷ GTF which is part of Bioproject PRJNA31257, The Human Genome Project, currently maintained by the Genome Reference Consortium (GRC) and the file is available for download at https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.26/. The RefSeq file was first filtered to select for CDS protein coding regions, then intersected with our neomer list for liver cancer to identify neomer-intersected genes in protein coding regions, using GenomicRanges in R package⁶⁹ (<https://bioconductor.org/packages/release/bioc/html/GenomicRanges.html>). Non-coding neomers were characterized for the following functional units: promoters, enhancers, and 3' UTR regions. Gene regulatory elements (distal enhancers, proximal enhancers, and promoters) were obtained from SCREEN V3 registry¹⁵, available for download at <https://screen.encodeproject.org/>. The hg38 knownGene 3' UTRs by transcript in R package⁷⁰ (<https://bioconductor.org/packages/release/data/annotation/html/TxDb.Hsapiens.UCSC.hg38.knownGene.html>) was used to annotate 3' UTRs.

Neomer gene target loci identification of noncoding regions in liver cancer

Neomer gene target loci annotation was carried out using The Open REGulatory ANNOtation database (OregAnno³⁰). We intersected our list of genomic noncoding neomers with the list of genomic regulatory elements using GenomicRanges package in R⁶⁹ to annotate our putative neomer gene target loci.

Gene Ontology

GO terms from biological processes were extracted from bioconductor species specific annotation packages (org.Hs.eg.db)⁷¹ and enrichGO from clusterProfiler³¹ in R. For non-coding regions, we first annotated the genomic loci of our neomers with experimentally validated genes using ORegAnno¹. We use this gene list stratified by enhancers, promoters and 3'UTRs using the following default parameters: Orgdb= "org.Hs.eg.db", keytype= "symbol", ontology=BP (biological process), p-value cutoff=0.05, adjusted p-value= BH (Benjamini-Hochberg). Qscore values were calculated by the -log base 10 of the adjusted p-value.

Patient recurrence analysis

To find patient recurrence, we use the Pandas library in Python⁷² to take the value counts of all neomers, grouped by: chromosome, start, stop, reference allele and alternate allele mutation. Next, we filter for neomers present in 3 or more patients. We further filtered neomers to remove mutations that create the same neomer in order to reduce cloning duplicates using a custom script. Here, we assume that each tumor sample is from a different patient.

1kb bin enrichment analysis

We created bins using the tileGenome function in the genomicRanges⁶⁹ package in R, with a tilewidth of 1000bp, with the option cut.last.tile.in.chrom=TRUE. We next intersected our noncoding neomer list with the aforementioned 1kb bins using the genomicRanges⁶⁹ package. We then filtered for regions that have 7 or more neomers per 1kb bin, defining these as regions of high neomer density or enriched neomer density.

Author contributions

J.S., O.Y.B., I.G.S., M.H., N.A. designed experiments. J.S., O.Y.B., C.D. performed MPRA experiments. J.S., I.G.S., R.E. generated data analysis, J.S., O.Y.B., I.G.S., M.H., W.L., J.L. contributed to study design. J.S. performed intersection analyses, genomic annotations, patient recurrence and 1kb neomer enrichment analysis.

Competing interests

N.A. is the cofounder and on the scientific advisory board of Regel Therapeutics and receives funding from BioMarin Pharmaceutical Incorporated. A patent application covering embodiments and concepts disclosed in the manuscript has been filed.

References

1. Choi, S. et al. Global burden of primary liver cancer and its association with underlying aetiologies, sociodemographic status, and sex differences from 1990-2019: A DALY-based analysis of the Global Burden of Disease 2019 study. *Clin. Mol. Hepatol.* **29**, 433–452 (2023).
2. Rao, C. V., Asch, A. S. & Yamada, H. Y. Frequently mutated genes/pathways and genomic instability as prevention targets in liver cancer. *Carcinogenesis* **38**, 2–11 (2017).
3. Khan, A. & Zhang, X. Function of the Long Noncoding RNAs in Hepatocellular Carcinoma: Classification, Molecular Mechanisms, and Significant Therapeutic Potentials. *Bioengineering (Basel)* **9**, (2022).
4. Kim, Y.-A., Park, K.-K. & Lee, S.-J. LncRNAs Act as a Link between Chronic Liver Disease and Hepatocellular Carcinoma. *Int. J. Mol. Sci.* **21**, (2020).
5. Luo, Y. et al. LncRNA SNHG17 Contributes to Proliferation, Migration, and Poor Prognosis of Hepatocellular Carcinoma. *Can J Gastroenterol Hepatol* **2021**, 9990338 (2021).
6. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
7. Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov.* **12**, 31–46 (2022).
8. Herz, H.-M., Hu, D. & Shilatifard, A. Enhancer malfunction in cancer. *Mol. Cell* **53**, 859–866 (2014).
9. Lowdon, R. F. & Wang, T. Epigenomic annotation of noncoding mutations identifies mutated pathways in primary liver cancer. *PLoS One* **12**, e0174032 (2017).
10. Tian, L. et al. Dysregulation of RalA signaling through dual regulatory mechanisms exerts its oncogenic functions in hepatocellular carcinoma. *Hepatology* **76**, 48–65 (2022).
11. Georgakopoulos-Soares, I. et al. Leveraging sequences missing from the human genome to

- diagnose cancer. *bioRxiv* (2021) doi:10.1101/2021.08.15.21261805.
12. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**, 159–164 (2015).
 13. Liu X, Xie W, Meng S, Kang X, Liu Y, Guo L, Wang C. Small Nucleolar RNAs and Their Comprehensive Biological Functions in Hepatocellular Carcinoma. *Cells*. **11**(17):2654 (2022).
 14. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
 15. ENCODE Project Consortium et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
 16. Wu, T. et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* **2**, 100141 (2021).
 17. Liu F, Liao Z, Zhang Z. MYC in liver cancer: mechanisms and targeted therapy opportunities. *Oncogene*. **42**(45):3303-3318 (2023).
 18. Liang J, Zhao H, Diplas BH, Liu S, Liu J, Wang D, Lu Y, Zhu Q, Wu J, Wang W, Yan H, Zeng YX, Wang X, Jiao Y. Genome-Wide CRISPR-Cas9 Screen Reveals Selective Vulnerability of ATRX-Mutant Cancers to WEE1 Inhibition. *Cancer Res*. **80**(3):510-523 (2020).
 19. Tian Y, Wu J, Chagas C, Du Y, Lyu H, He Y, Qi S, Peng Y, Hu J. CDCA5 overexpression is an Indicator of poor prognosis in patients with hepatocellular carcinoma (HCC). *BMC Cancer*. **18**(1):1187 (2018).
 20. Kong DG, Yao FZ. CDC6 is a possible biomarker for hepatocellular carcinoma. *Int J Clin Exp Pathol*. **14**(7):811-818 (2021).
 21. Ding Y, Sun Z, Zhang S, Han X, Li Y, Xu Q, Zhou L, Xu H, Bai Y, Xu C, Ding H, Ge Y, Wang W. Down-regulation of small nuclear RNA (snRNA) RNU5E-1 in hepatocellular carcinoma

- presents with vital clinical significance. *J Gastrointest Oncol.* **11**(4):738-746 (2020).
22. Zhou R, Chen J, Xu Y, Ye Y, Zhong G, Chen T, Qiu L. PRPF19 facilitates colorectal cancer liver metastasis through activation of the Src-YAP1 pathway via K63-linked ubiquitination of MYL9. *Cell Death Dis.* **14**(4):258 (2023).
 23. Knudsen ES, Nambiar R, Rosario SR, Smiraglia DJ, Goodrich DW, Witkiewicz AK. Pan-cancer molecular analysis of the RB tumor suppressor pathway. *Commun Biol.* **3**(1):158 (2020).
 24. Zhao H, Wang Y, Yang Z, Wei W, Cong Z, Xie Y. High expression of aurora kinase B predicts poor prognosis in hepatocellular carcinoma after curative surgery and its effects on the tumor microenvironment. *Ann Transl Med.* **10**(21):1168 (2022).
 25. Zhang Y, Liang B, Song X, Wang H, Evert M, Zhou Y, Calvisi DF, Tang L, Chen X. Loss of Apc Cooperates with Activated Oncogenes to Induce Liver Tumor Formation in Mice. *Am J Pathol.* **191**(5):930-946 (2021).
 26. Huang WC, Tung SL, Chen YL, Chen PM, Chu PY. IFI44L is a novel tumor suppressor in human hepatocellular carcinoma affecting cancer stemness, metastasis, and drug resistance via regulating met/Src signaling pathway. *BMC Cancer.* **18**(1):609 (2018).
 27. Prasad, M., Brzostek, J., Gautam, N. et al. Themis regulates metabolic signaling and effector functions in CD4⁺ T cells by controlling NFAT nuclear translocation. *Cell Mol Immunol* **18**, 2249–2261 (2021).
 28. He FY, Chen G, He RQ, Huang ZG, Li JD, Wu WZ, Chen JT, Tang YL, Li DM, Pan SL, Feng ZB, Dang YW. Expression of IER3 in hepatocellular carcinoma: clinicopathology, prognosis, and potential regulatory pathways. *PeerJ.* **10**:e12944 (2022).
 29. Paturel A, Hall J, Chemin I. Poly(ADP-Ribose) Polymerase Inhibition as a Promising Approach for Hepatocellular Carcinoma Therapy. *Cancers (Basel).* **14**(15):3806 (2022).

30. Lesurf, R. et al. ORegAnno 3.0: a community-driven resource for curated regulatory annotation. *Nucleic Acids Res.* **44**, D126–32 (2016).
31. Wu, T. et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* **2**, 100141 (2021).
32. van Zijl, F. et al. Epithelial-mesenchymal transition in hepatocellular carcinoma. *Future Oncol.* **5**, 1169–1179 (2009).
33. Zhang, Yi-Gan, et al. “Reclassification of Hepatocellular Cancer with Neural-Related Genes.” *Frontiers in Oncology* (2022).
34. Dropmann, A. et al. TGF- β 1 and TGF- β 2 abundance in liver diseases of mice and men. *Oncotarget* **7**, 19499–19518 (2016).
35. Wang, T. et al. RHO GTPase family in hepatocellular carcinoma. *Exp. Hematol. Oncol.* **11**, 91 (2022).
36. Blanc, L. et al. Critical function for the Ras-GTPase activating protein RASA3 in vertebrate erythropoiesis and megakaryopoiesis. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 12099–12104 (2012).
37. Lin, H., Fan, X., He, L. & Zhou, D. Methylation patterns of RASA3 associated with clinicopathological factors in hepatocellular carcinoma. *J. Cancer* **9**, 2116–2122 (2018).
38. Li, Y., Yin, Y., He, Y., He, K. & Li, J. SOS1 regulates HCC cell epithelial-mesenchymal transition via the PI3K/AKT/mTOR pathway. *Biochem. Biophys. Res. Commun.* **637**, 161–169 (2022).
39. Liu, A.-W. et al. ShRNA-targeted MAP4K4 inhibits hepatocellular carcinoma growth. *Clin. Cancer Res.* **17**, 710–720 (2011).
40. Wang, W., Smits, R., Hao, H. & He, C. Wnt/ β -Catenin Signaling in Liver Cancers. *Cancers* **11**, (2019).

41. Lan, L., Wang, W., Huang, Y., Zhao, C. & Bu, X. WNT7A Overexpression Inhibits Growth and Migration of Hepatocellular Carcinoma via the β -Catenin Independent Pathway. *Biomed Res. Int.* **2019**, 3605950 (2019).
42. Behari, J. The Wnt/ β -catenin signaling pathway in liver biology and disease. *Expert Rev. Gastroenterol. Hepatol.* **4**, 745–756 (2010).
43. Liu Y, Yang C, Chen S, Liu W, Liang J, He S, Hui J. Cancer-derived exosomal miR-375 targets DIP2C and promotes osteoblastic metastasis and prostate cancer progression by regulating the Wnt signaling pathway. *Cancer Gene Ther.* **30**(3):437-449 (2023).
44. Liang X, Du L, Fan Y. The potential of FCRL genes as targets for cancer treatment: insights from bioinformatics and immunology. *Aging (Albany NY).* **15**(11):4926-4948 (2023).
45. Chen L, Shi H, Wang X, Wang T, Wang Y, Wu Z, Zhang W, Chen H, Zhong M, Mao X, Shi X, Li Q. Hepatocyte nuclear factor 4 gamma (HNF4G) is correlated with poor prognosis and promotes tumor cell growth by inhibiting caspase-dependent intrinsic apoptosis in colorectal cancer. *Eur J Pharmacol.* **916**:174727 (2022).
46. Guo X, Yang M, Gu H, Zhao J, Zou L. Decreased expression of SOX6 confers a poor prognosis in hepatocellular carcinoma. *Cancer Epidemiol.* **37**(5):732-6 (2013).
47. Yang Z, Jiang S, Lu C, Ji T, Yang W, Li T, Lv J, Hu W, Yang Y, Jin Z. SOX11: friend or foe in tumor prevention and carcinogenesis? *Ther Adv Med Oncol.* **11**:1758835919853449 (2019).
48. Joo JS, Cho SY, Rou WS, Kim JS, Kang SH, Lee ES, Moon HS, Kim SH, Sung JK, Kwon IS, Eun HS, Lee BS. TEAD2 as a novel prognostic factor for hepatocellular carcinoma. *Oncol Rep.* **43**(6):1785-1796 (2020).
49. Horie Y, Suzuki A, Kataoka E, Sasaki T, Hamada K, Sasaki J, Mizuno K, Hasegawa G, Kishimoto H, Iizuka M, Naito M, Enomoto K, Watanabe S, Mak TW, Nakano T. Hepatocyte-

- specific Pten deficiency results in steatohepatitis and hepatocellular carcinomas. *J Clin Invest.* **113**(12):1774-83 (2004).
50. Rao S, Mishra L. Targeting Transforming Growth Factor Beta Signaling in Liver Cancer. *Hepatology.* **69**(4):1375-1378 (2019).
51. Fang Y, Cao H, Gong X, Chen Y, Zhuang Y, Zhou S, Chen Y, Jiang Y, Ji X, Peng H, Jing X. AFF4 Predicts the Prognosis of Colorectal Cancer Patients and Suppresses Colorectal Cancer Metastasis via Promoting CDH1 Expression. *Front Oncol.* **12**:797392 (2022).
52. Yuan, J., Ju, Q., Zhu, J. et al. RASSF9 promotes NSCLC cell proliferation by activating the MEK/ERK axis. *Cell Death Discov.* **7**, 199 (2021).
53. Liu W, Wang J, Wang L, Qian C, Qian Y, Xuan H, Zhuo W, Li X, Yu J, Si J. Ras-association domain family 10 acts as a novel tumor suppressor through modulating MMP2 in hepatocarcinoma. *Oncogenesis.* **5**(6):e237 (2016).
54. Liu, J. et al. Identification of liver metastasis-associated genes in human colon carcinoma by mRNA profiling. *Chin. J. Cancer Res.* **30**, 633–646 (2018).
55. Cui, X.-X. et al. High expression of ZNF93 promotes proliferation and migration of ovarian cancer cells and relates to poor prognosis. *Int. J. Clin. Exp. Pathol.* **13**, 944–953 (2020).
56. Zhang, Haibin, et al. “Hepatocellular Carcinoma Carrying TERT Promoter Mutations: JHC.” *Journal of Hepatocellular Carcinoma*, Dove Press (2022).
57. Farman, F. U., Iqbal, M., Azam, M. & Saeed, M. Nucleosomes positioning around transcriptional start site of tumor suppressor (Rbl2/p130) gene in breast cancer. *Mol. Biol. Rep.* **45**, 185–194 (2018).
58. Wang Y, Liu D, Zhang T, Xia L. FGF/FGFR Signaling in Hepatocellular Carcinoma: From Carcinogenesis to Recent Therapeutic Intervention. *Cancers (Basel).* **13**(6):1360 (2021).

59. Yuan, J., Chen, L., Xiao, J. et al. SHROOM2 inhibits tumor metastasis through RhoA–ROCK pathway-dependent and -independent mechanisms in nasopharyngeal carcinoma. *Cell Death Dis* **10**, 58 (2019).
60. Wang, T., Rao, D., Yu, C. et al. RHO GTPase family in hepatocellular carcinoma. *Exp Hematol Oncol* **11**, 91 (2022).
61. Chew, T., Liu, X., Liu, L. et al. Crosstalk of Ras and Rho: activation of RhoA abates Kras-induced liver tumorigenesis in transgenic zebrafish models. *Oncogene* **33**, 2717–2727 (2014).
62. Patergnani, S., Buchsbaum, D. & Piazza, G. Targeting the Wnt/ β -catenin Signaling Pathway in Cancer. (Frontiers Media SA, 2022).
63. Zhou X, Zheng Y. Cell type-specific signaling function of RhoA GTPase: lessons from mouse gene targeting. *J Biol Chem.* **288**(51):36179-88 (2013).
64. Wang, T., Rao, D., Yu, C. et al. RHO GTPase family in hepatocellular carcinoma. *Exp Hematol Oncol* **11**, 91 (2022).
65. Chew TW, Liu XJ, Liu L, Spitsbergen JM, Gong Z, Low BC. Crosstalk of Ras and Rho: activation of RhoA abates Kras-induced liver tumorigenesis in transgenic zebrafish models. *Oncogene.* **33**(21):2717-27 (2014).
66. Bai Y, Xie F, Miao F, Long J, Huang S, Huang H, Lin J, Wang D, Yang X, Bian J, Mao J, Wang X, Mao Y, Sang X, Zhao H. The diagnostic and prognostic role of RhoA in hepatocellular carcinoma. *Aging (Albany NY).* **11**(14):5158-5172 (2019).
67. O'Leary NA, Wright MW, Brister JR, Murphy TD, Pruitt KD, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**(D1):D733-45 (2016).
68. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human

genome browser at UCSC. *Genome Res.* **12**(6):996-1006 (2002).

69. Lawrence, M. et al. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
70. Team BC, Maintainer BP. `TxDb.Hsapiens.UCSC.hg38.knownGene`: Annotation package for TxDb object(s). R package version 3.4.6. (2019).
71. Carlson, M., Falcon, S., Pages, H. & Li, N. org. `Hs. eg. db`: Genome wide annotation for Human. R package version.
72. “Pandas.” PyPI, pypi.org/project/pandas/.

Publishing Agreement

It is the policy of the University to encourage open access and broad distribution of all theses, dissertations, and manuscripts. The Graduate Division will facilitate the distribution of UCSF theses, dissertations, and manuscripts to the UCSF Library for open access and distribution. UCSF will make such theses, dissertations, and manuscripts accessible to the public and will take reasonable steps to preserve these works in perpetuity.

I hereby grant the non-exclusive, perpetual right to The Regents of the University of California to reproduce, publicly display, distribute, preserve, and publish copies of my thesis, dissertation, or manuscript in any form or media, now existing or later derived, including access online for teaching, research, and public service purposes.

DocuSigned by:
Jasmine Sims
F1AA36A6C795479...

Author Signature

3/21/2024

Date