# Manuscript revision note

Dear Editor,

Greetings! We sincerely appreciate the valuable feedback provided by you and the esteemed reviewers on our manuscript (Manuscript ID: JTRM-D-23-06892R1, "Machine learning and multi-omics data reveal driver gene-based molecular subtypes in hepatocellular carcinoma for precision treatment"). We have thoroughly reviewed all the comments and are immensely grateful for the insightful suggestions made by you and the reviewers. Subsequently, we have diligently revised the manuscript based on your recommendations. We earnestly hope that these revisions will make our manuscript suitable for publication in the "*Journal of Translational Medicine*." We firmly believe that the refined manuscript will better resonate with the readers of the journal.

General revisions made to the manuscript include:

1. Highlighting all modifications in yellow throughout the manuscript.

2. Addition of a new reference, "Hanahan D: Hallmarks of Cancer: New Dimensions. Cancer Discov 2022, 12:31-46," and corresponding adjustments in the text, resulting in changes to the reference numbering.

3. As suggested by the reviewers, the incorporation of a supplementary table, Supplementary Table S5, delineates the details of the cohorts used in the study. This has led to a renumbering of the supplementary tables; for instance, the original Supplementary Table S5 is now Supplementary Table S7.

4. While responding to the reviewers' comments, we attempted to include images for better elucidation. However, we noticed that Editorial Manager does not support the upload of images or tables. Consequently, we have uploaded these response contents with images to GitHub for the reviewers' reference.

5. We have provided a point-by-point response to the reviewers' comments, hoping that these revisions meet your and the reviewers' expectations.

Once again, we extend our gratitude for your and the reviewers' patient guidance and invaluable suggestions.

Best regards,

Xiaoqin Li

Beijing University of Technology

lxq0811@bjut.edu.cn

## Reviewer #1

1.Sample information. The authors use 3 HCC datasets in this study. However, they don't include a detailed description of the cohorts. They must include a table or tables with number of samples in each cohort, groups, clinical information, number of genes, etc.

**Response**:

Thank you for your review and valuable suggestions regarding our study. We have addressed your recommendation by including an additional table in the supplementary materials, providing a detailed description of sample information, clinical characteristics, gene counts, and other relevant details for each dataset（Supplementary Table S5）. We genuinely appreciate your valuable input, which has significantly contributed to the improvement of our research

Supplementary Table S5 Description of the patient cohort

| Clinical | Characteristics | TCGA_LIHC（n = 357） | ICGC_JP(n = 226) | ICGC_FR(n = 107) |
|---|---|---|---|---|
| Age | Median [1st Qu,3rd Qu] | 59 [52，68] | 67 [62，74] | 64 [57，73] |
| Gender | Male | 241 | 167 | 83 |
| | Female | 116 | 59 | 24 |
| Stage | I | 166 | 35 | NA |
| | II | 84 | 104 | NA |
| | III | 80 | 69 | NA |
| | IV | 5 | 18 | NA |
| | NA | 22 | 0 | NA |
| Grade | G1 | 51 | NA | NA |
| | G2 | 174 | NA | NA |
| | G3 | 116 | NA | NA |
| | G4 | 11 | NA | NA |
| | NA | 5 | NA | NA |
| Alcohol consumption | Yes | 113 | NA | NA |
| | No | 226 | NA | NA |
| | NA | 18 | NA | NA |
| HBV | Yes | 139 | NA | NA |
| | No | 218 | NA | NA |
| HCV | Yes | 99 | NA | NA |
| | No | 258 | NA | NA |
| Vascular invasion | Macro | 14 | NA | NA |
| | Micro | 89 | NA | NA |
| | None | 199 | NA | NA |
| | NA | 55 | NA | NA |
| AFP(ng/mL) | Median[1st Qu,3rd Qu] | 14251 [4, 282] | NA | NA |
| Platform | | Illumina HiSeq 2000 | Illumina HiSeq | Illumina HiSeq 2000 |

| | | | |
|---|---|---|---|
| | platform | 2000 platform | platform |
| Number of RNA-seq Gene | 60499 | 22913 | 557820 |
| Country | USA | Japan | France |

2. Also, for RNAseq, how many genes remained after removing genes. Why did you use TPM values<0 in more than 30% of samples?

**Response**:

We gratefully appreciate your valuable comment. After removing lowly expressed genes, approximately 17,660 genes were retained for subsequent analysis. We have made the necessary clarification in the methods section of the manuscript. Additionally, we'd like to acknowledge a typographical error in our previous response. We removed genes with TPM values ≤ 0 in more than 30% of samples. In RNA-Seq data analysis, TPM values ≤ 0 are often attributed to technical errors or sequencing noise. This approach is a common method to ensure data quality and reduce noise and has been applied in previous studies (e.g., Jiang et al[1]. and Li et al[2].).

Reference:

[1] Jiang, Yi-Zhou, et al. "Genomic and transcriptomic landscape of triple-negative breast cancers: subtypes and treatment strategies." Cancer cell 35.3 (2019): 428-440.

[2] Li, Qian, et al. "lncDIFF: a novel quasi-likelihood method for differential expression analysis of non-coding RNA." BMC genomics 20 (2019): 1-13.

3.What criteria did you use to define the initial 96 driver genes? Could you add the rationale behind using these 3 specific studies and the selection method?

**Response**:

We appreciate your question regarding the criteria for defining the initial 96 driver genes and the rationale for choosing these specific studies. We provide a supplementary explanation of this issue in Results Section 3.1. Here is our response:

First and foremost, we selected these three studies to obtain a more comprehensive and widely recognized list of HCC driver genes.

First, Bailey et al.'s[1] study, as part of the TCGA Pan-Cancer Atlas, employed a driver gene discovery algorithm based on single-gene mutation frequencies and a manual curation process to identify 29 driver genes. This study is widely regarded as a seminal work in the field of driver genes because it integrates data from multiple cancer types and provides a comprehensive gene list.

Second, the study by Martínez-Jiménez et al.[2], published in 2020, offers the most comprehensive catalogue of driver genes in cancer to date. While they did not involve experts in the curation of results, compared to Bailey et al.'s study, Martínez-Jiménez et al. analyzed a larger number of samples. They employed a more comprehensive approach to driver gene exploration, thus providing valuable insights.

Finally, Fujimoto et al.'s[3] study focused specifically on HCC, integrating the analysis of point mutations, structural variations, and viral integrations in both coding and non-coding regions of 300 HCC samples to identify HCC driver genes based on mutation frequencies. This study contributes to a better understanding of specific driver genes in HCC.

By selecting these three studies, we aimed to obtain a more comprehensive and widely recognized list of HCC driver genes to enhance our understanding of the driver gene mechanisms in HCC.

Reference:

[1] Bailey, Matthew H., et al. "Comprehensive characterization of cancer driver genes and mutations." Cell 173.2 (2018): 371-385.

[2] Martínez-Jiménez, Francisco, et al. "A compendium of mutational cancer driver genes." Nature Reviews Cancer 20.10 (2020): 555-572.

[3] Fujimoto, Akihiro, et al. "Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer." Nature genetics 48.5 (2016): 500-509.

4.How did you validate the gene family expansion method? How did they ensure that the expanded driver genes are relevant and specific to HCC?

**Response**:

Thank you for your valuable suggestions. In fact, we have separately written a paper on the issues you mentioned, but unfortunately, it is still in the process of being published. Here are some preliminary results from our study on the extended list of driver genes:

After an in-depth exploration of the extended list of driver genes, we found that in humans, transcription factors constitute 8% of all genes and are associated with various diseases and phenotypes, with their mutations often carrying severe deleterious effects [1]. Previous studies have indicated frequent mutations in the zinc finger transcription factor family in endometrial cancer, colorectal cancer, and melanoma, contributing to widespread transcriptional dysregulation observed in cancer cells[2]. In our research, we observed an enrichment of mutations in the ZF-H2C2_2 domain in HCC, with nine mutation hotspots. The spatial structure of zf-H2C2_2 consists of an αββ structure formed by two consecutive C2H2-type zinc finger structures (Figure 1B). Through conservation analysis of mutation hotspots, we found that the majority of these hotspots are highly conserved (Figure 1A). Among them, the fifth and tenth residues of histidine are one of the four residues coordinating with zinc (Figure 1C), crucial for the stability of the domain, and substitutions in other residues may lead to the loss of domain function [3]. According to Barrera et al.'s study [4], we found that the residue at position 35 is connected to the DNA backbone, and its mutation may alter the binding affinity or specificity with DNA. Additionally, experimental studies have identified some zinc finger transcription factors closely associated with cancer progression. For example, ectopic expression of *ZNF331* reduces the migration and invasion capabilities of cancer cells[5]. Mutations or translocations of *ZNF687* are closely associated with the occurrence and development of cancer[6-7]. In HCC, the overexpression of *ZNF687* significantly promotes liver cancer stem cell-like characteristics and tumour formation by upregulating the transcription levels of pluripotency-related factors BMI1, OCT4, and NANOG[8]. Mutations in the zinc finger domain of *HNF4A* often result in transcriptional dysregulation in HCC, promoting tumour growth[9]. The mutation frequency of these genes is extremely low (sometimes even less than 1%), making them challenging to identify in traditional single-gene driver gene identification algorithms. Martinez-Jimenez et al.'s study on a cohort of 1616 HCC samples could confidently identify only 1-2 members. The use of gene family extension can be considered a form of data augmentation, expanding the concept of repeated changes in individual genes in cancer to family domains with similar structures,

thereby enhancing the statistical performance of the algorithm and uncovering more potential driver genes.

The main purpose of this manuscript is to classify HCC, explore the biological differences among different subtypes, and discover clinical diagnostic markers. We believe that including the validation of driver genes would make the entire manuscript excessively lengthy. Therefore, in the discussion section of this paper, we provided only a general description of this aspect. For example, we found that many significantly mutated domains are related to classical oncogenic signalling pathways, such as genes involved in the RTK signalling pathway, including the pkase_ty and SH2 domains. Downstream in the PI3K/AKT signaling pathway, domains such as PI3_PI4_kinase, PH, and PI3Ka are also enriched with mutations. Similarly, we observed enrichment of mutations in the SET domain, a conserved catalytic core in the histone lysine methyltransferase family, crucial for the tumour-suppressive function according to [10].

Reference:


[1] Lambert, Samuel A., et al. "The human transcription factors." Cell 172.4 (2018): 650-665.

[2] Munro, Daniel, Dario Ghersi, and Mona Singh. "Two critical positions in zinc finger domains are heavily mutated in three human cancer types." PLoS computational biology 14.6 (2018): e1006290.

[3] Wolfe, Scot A., Lena Nekludova, and Carl O. Pabo. "DNA recognition by Cys2His2 zinc finger proteins." Annual review of biophysics and biomolecular structure 29.1 (2000): 183-212.

[4] Barrera, Luis A., et al. "Survey of variation in human transcription factors reveals prevalent DNA binding changes." Science 351.6280 (2016): 1450-1454.

[5] Yu, J., et al. "Zinc-finger protein 331, a novel putative tumor suppressor, suppresses growth and invasiveness of gastric cancer." Oncogene 32.3 (2013): 307-317.

[6] Divisato, Giuseppina, et al. "ZNF687 mutations in severe Paget disease of bone associated with giant cell tumor." The American Journal of Human Genetics 98.2 (2016): 275-286.

[7] Nguyen, TuDung T., et al. "Identification of novel Runx1 (AML1) translocation partner genes SH3D19, YTHDf2, and ZNF687 in acute myeloid leukemia." Genes, Chromosomes and Cancer 45.10 (2006): 918-932.

[8] Zhang, T., et al. "Overexpression of zinc finger protein 687 enhances tumorigenic capability and promotes recurrence of hepatocellular carcinoma." Oncogenesis 6.7 (2017): e363-e363.

[9] Taniguchi, Hiroaki, et al. "Loss-of-function mutations in Zn-finger DNA-binding domain of HNF4A cause aberrant transcriptional regulation in liver cancer." Oncotarget 9.40 (2018): 26144.

[10] Kim, Keun-Cheol, Liqing Geng, and Shi Huang. "Inactivation of a histone methyltransferase by mutations in human cancers." Cancer research 63.22 (2003): 7619-7623.


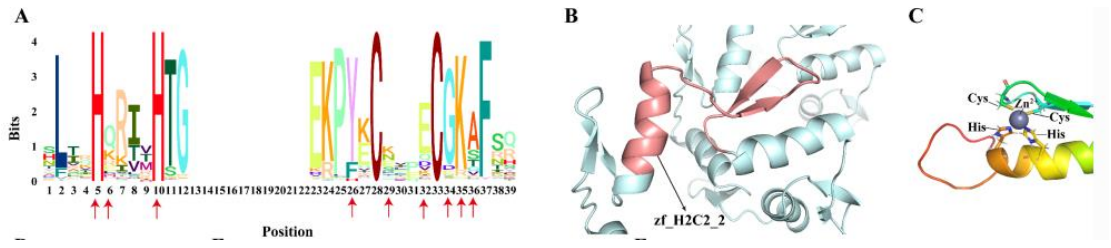


Fig. 1. Mutation hotspots in zf-H2C2_2 (A) Sequence logo of zf-H2C2_2 alignment(only shown

position with significant mutations), red arrows indicate mutation hotspots (B) Spatial structure of zf-H2C2_2 (This structure is from Alpha Fold, and the low confidence region is removed) (C) Spatial structure of C2H2 zinc finger domains(PDB：2eoz)

| 5. In point 3.5 and S7, the authors find statistical significance in clinical characteristics between Class A and B. Could you include a discussion on how this is affecting the subtypes separation? How much do you think that this multiomics analysis contributes? |
|---|

**Response**:

We appreciate your feedback. In Sections 3.5 and S7, we did indeed find statistically significant differences in clinical characteristics between Class A and Class B patients. These differences reflect the heterogeneity between different subtypes, which is a common phenomenon in cancer research. However, we recognize that these clinical features' differences (such as age and tumour stage) may themselves be related to patient survival, making it challenging to determine whether the observed survival differences between subtypes are due to these clinical characteristics or whether the subtypes have independent prognostic factors. Therefore, to gain a more comprehensive understanding of the differences between Class A and Class B subtypes, we introduced univariate and multivariate Cox regression models. The analysis from these models confirmed that the subtypes we defined are independent factors in the survival prognosis of HCC patients, thereby eliminating the potential impact of clinical features.

Furthermore, the multi-omics features of the subtypes may also affect their distinct clinical characteristics. For example, as mentioned in our discussion section, significant chromosomal instability was observed in the Class B subtype, along with deletions in chromosomes 4p, 4q, 13q, 16p, 16q, and 17p, which may result in more clinically advanced and poorly differentiated HCC patients in the Class B subtype[1-2].

We believe that this comprehensive analysis of multi-omics and clinical characteristics contributes to a better understanding of survival differences between different subtypes and provides essential information for personalized treatment.

Reference

[1] Tsuda, Hitoshi, et al. "Allele loss on chromosome 16 associated with progression of human hepatocellular carcinoma." Proceedings of the National Academy of Sciences 87.17 (1990): 6791-6794.

[2] Nishida, Naoshi, et al. "Accumulation of allelic loss on arms of chromosomes 13q, 16q and 17p in the advanced stages of human hepatocellular carcinoma." International journal of cancer 51.6 (1992): 862-868.

| 6. Also, can you implement the rationale behind the microbiome analysis? Population? How did you compare with your data? |
|---|

**Response**:

We appreciate your suggestion and in our discussion section, we have provided the rationale for analyzing microbiome data in our study. The polymorphic microbiome has been recognized for its association with prominent cancer hallmark such as sustaining proliferative signaling, resisting cell death, and enabling replicative immortality[1]. Several studies have found a significant link between a polymorphic microbiome and cancer progression. For example, Jin et al.[2] identified

that the local microbiota in lung cancer triggers inflammation associated with lung adenocarcinoma by activating γδ T cells within the lung. Mice treated with sterile conditions or antibiotics showed a marked reduction in lung cancer development caused by Kras mutation and p53 deletion. Similarly, Bullman et al.[2] revealed a close correlation between Fusobacterium and the progression of colorectal cancer, with antibiotic treatment reducing the Fusobacterium load and inhibiting tumor growth in colon cancer xenograft mice.

In HCC, the gut microbiota influences the liver through metabolites, subsequently impacting immune cell functionality[4]. In 2020, Poore et al.[5] assessed microbial abundance in TCGA cancer cohorts by analyzing genomic sequencing and RNA-seq raw data. They emphasized the distinct microbial compositions in each cancer type and substantial differences between cancerous and normal samples. However, these differences were not significant within the same cancer type, such as between tumors at different stages. Given the unavailability of microbiome data in our study cohort, we leveraged Poore et al.'s results to explore differences in microbial characteristics among subtypes. While Poore et al.'s findings cannot replace the clinical gold standard for microbiome detection, our results suggest a trend – that microbial features significantly differ among subtypes of the same cancer. This trend may hold potential for diagnostic and therapeutic applications in cancer, though further exploration by experts in the field is warranted.

Reference:

[1] Hanahan, Douglas. "Hallmarks of cancer: new dimensions." Cancer discovery 12.1 (2022): 31-46.

[2] Jin C, Lagoudas G K, Zhao C, et al. Commensal microbiota promote lung cancer development via γδ T cells[J]. Cell, 2019, 176(5): 998-1013. e16.

[3] Bullman, Susan, et al. "Analysis of Fusobacterium persistence and antibiotic response in colorectal cancer." Science 358.6369 (2017): 1443-1448.

[4] Ma C, Han M, Heinrich B, et al. Gut microbiome–mediated bile acid metabolism regulates liver cancer via NKT cells[J]. Science, 2018, 360(6391): eaan5931.

[5] Poore, Gregory D., et al. "Microbiome analyses of blood and tissues suggest cancer diagnostic approach." Nature 579.7800 (2020): 567-574.

7. Did you compare your models with existing methods or tools? Could you validate your findings using another method?

**Response**:

In our study, we employed an unconventional approach utilizing both mutation and gene expression data to define molecular subtypes of HCC. This methodology is not commonly utilized in existing cancer stratification tools. Typically, mutation data is binary (0 or 1), making it unsuitable for measuring similarity between patients using Euclidean distances or correlation coefficients as with continuous data. This does not align with the requirements of many current cancer stratification tools. To our knowledge, only the iCluster tool supports the input of data containing binary discrete variables. While iCluster exhibited statistical significance in our dataset, the subtypes it generated did not meet our expectations. In contrast, our proposed approach resulted in more significant clinical subtypes (0.0011 vs 7.88e-05), better aligning with our research findings

To validate the effectiveness of our subtype classification, we compared it to previously published HCC subtypes. This section was originally in Results 3.5, but we moved it to section 3.2 for coherence. Specifically, we compared our subtypes to those defined by Hosdia[1], Benfeitas[2], Bidkhori[3], and TCGA[4], among others, and found a significant correlation between them (see Supplementary Figure S8). CLASS A subtypes are enriched in several high-metabolism subtypes, which are generally associated with better prognosis, such as Hosdia S3, iHCC1, and hALDH2. Conversely, CLASS B subtypes are enriched in subtypes characterized by lower metabolic flux, a high mutation rate in the TP53 gene, and high cell proliferation levels, such as Hosdia S1, Hosdia S2, and iHCC3.

Additionally, based on the results of Benfeitas et al[3].'s study, we also discovered that these two subtypes may employ different biological mechanisms to counteract the effects of reactive oxygen species (ROS). This suggests the reliability and clinical relevance of the HCC subtypes we have defined.

Reference:

[1] Hoshida Y, Nijman SM, Kobayashi M, Chan JA, Brunet JP, Chiang DY, Villanueva A, Newell P, Ikeda K, Hashimoto M, et al: Integrative transcriptome analysis reveals common molecular subclasses of human hepatocellular carcinoma. Cancer Res 2009, 69:7385-7392.

[2] Benfeitas R, Bidkhori G, Mukhopadhyay B, Klevstig M, Arif M, Zhang C, Lee S, Cinar R, Nielsen J, Uhlen M, et al: Characterization of heterogeneous redox responses in hepatocellular carcinoma patients using network analysis. EBioMedicine 2019, 40:471-487.

[3] Bidkhori G, Benfeitas R, Klevstig M, Zhang C, Nielsen J, Uhlen M, Boren J, Mardinoglu A: Metabolic network-based stratification of hepatocellular carcinoma reveals three distinct tumor subtypes. Proc Natl Acad Sci U S A 2018, 115:E11874-E11883.

[4] Cancer Genome Atlas Research Network. Electronic address wbe, Cancer Genome Atlas Research N: Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell 2017, 169:1327-1341 e1323.

8. In general, this article relies heavily on theoretical concepts and needs to improve the rationale behind it. Also, they don't develop any new tools and rely on already existing ones. By itself, the idea and the analysis seem correct, but including experimental data and using other analytical methods could strengthen its validity and practical significance. Also, a rationale about the strength of the methods used is incomplete.

Also, the study lacks direct clinical applications. Real-world implications and potential therapeutic strategies could be better emphasized.

**Response**:

Thank you for your thorough review of our study and valuable suggestions. Indeed, our research is built upon a series of theoretical concepts. In response to your previous queries, we have provided detailed explanations in the manuscript regarding the methodology and related concepts, encompassing the description of the cohorts, RNA-seq data processing methods, and the selection of driver genes. We believe that these modifications will strengthen the theoretical foundation

behind the article.

Regarding the second point, although we have not introduced a completely new algorithm, we proposed an innovative computational framework that integrates existing tools, resulting in a more robust analytical process. In our study, this framework allows for the categorization of patients into subtypes utilizing mutation data and other omics data. In comparison to iCluster, another tool that also accepts mutation data, our method accurately defines HCC subtypes, reflecting more significant clinical implications. Additionally, our classification method is approximately 16 times faster than iCluster and can be operated on a standard home PC with a minimum of 16GB of RAM.

Regarding the third point, indeed, due to limitations in our experimental conditions, we did not conduct direct biological validations. To address this limitation, we have compared most of our findings with related biological experimental results, finding consistency between our data mining results and existing biological experimental studies. For example, the ten identified classifier genes we discovered all contribute to HCC progression[1-7], and the effect of *TTK* and its inhibitor on HCC is consistent with the referenced papers. The sensitivity of five cell lines (Hep3B, Huh7, MHCC97L, PLC/PRF/5, and Hepa1-6) to CFI-402257 (a TTK inhibitor)[8], with most (Hep3B, Huh7, and PLC/PRF/5) considered as CLASS B cell lines in our analysis, further supports our conclusions (Figure 9A). In general, nearly 80 of the references cited in this paper are mostly based on biological experiments, providing collateral evidence supporting the reliability and practical significance of our conclusions.

Finally, the direct application of our research findings to clinical settings is challenging due to the heterogeneity and diversity in omics data. Moving forward, we intend to continually monitor publicly available datasets to further collect data that meet our criteria and validate our results. Despite this, we have found a strong correlation between our subtype results and previously published ones, suggesting shared molecular characteristics among HCC subtypes identified by different methods. To reach a consensus among different subtype methods, we plan to conduct further research in the future.

Reference:

[1] Dai B, Zhang X, Shang R, Wang J, Yang X, Zhang H, Liu Q, Wang D, Wang L, Dou K: Blockade of ARHGAP11A reverses malignant progress via inactivating Rac1B in hepatocellular carcinoma. Cell Commun Signal 2018, 16:99.

[2] Zhang Z, Zhang Y, Mo W: The Autophagy Related Gene CHAF1B Is a Relevant Prognostic and Diagnostic Biomarker in Hepatocellular Carcinoma. Front Oncol 2020, 10:626175.

[3] Dang XW, Pan Q, Lin ZH, Wang HH, Li LH, Li L, Shen DQ, Wang PJ: Overexpressed DEPDC1B contributes to the progression of hepatocellular carcinoma by CDK1. Aging (Albany NY) 2021, 13:20094-20115.

[4] Chen J, Xia H, Zhang X, Karthik S, Pratap SV, Ooi LL, Hong W, Hui KM: ECT2 regulates the Rho/ERK signalling axis to promote early recurrence in human hepatocellular carcinoma. J Hepatol 2015, 62:1287-1295.

[5] Li S, Wu L, Zhang H, Liu X, Wang Z, Dong B, Cao G: GINS1 Induced Sorafenib Resistance by Promoting Cancer Stem Properties in Human Hepatocellular Cancer Cells. Front Cell Dev Biol 2021, 9:711894.

[6] Wu X, Wang H, Lian Y, Chen L, Gu L, Wang J, Huang Y, Deng M, Gao Z, Huang Y: GTSE1 promotes cell migration and invasion by regulating EMT in hepatocellular carcinoma and

is associated with poor prognosis. Sci Rep 2017, 7:5129.

[7] Yang Y, Gao L, Chen J, Xiao W, Liu R, Kan H: Lamin B1 is a potential therapeutic target and prognostic biomarker for hepatocellular carcinoma. Bioengineered 2022, 13:9211-9231

[8] Chan, Cerise Yuen-Ki, et al. "CFI-402257, a TTK inhibitor, effectively suppresses hepatocellular carcinoma." Proceedings of the National Academy of Sciences 119.32 (2022): e2119514119.

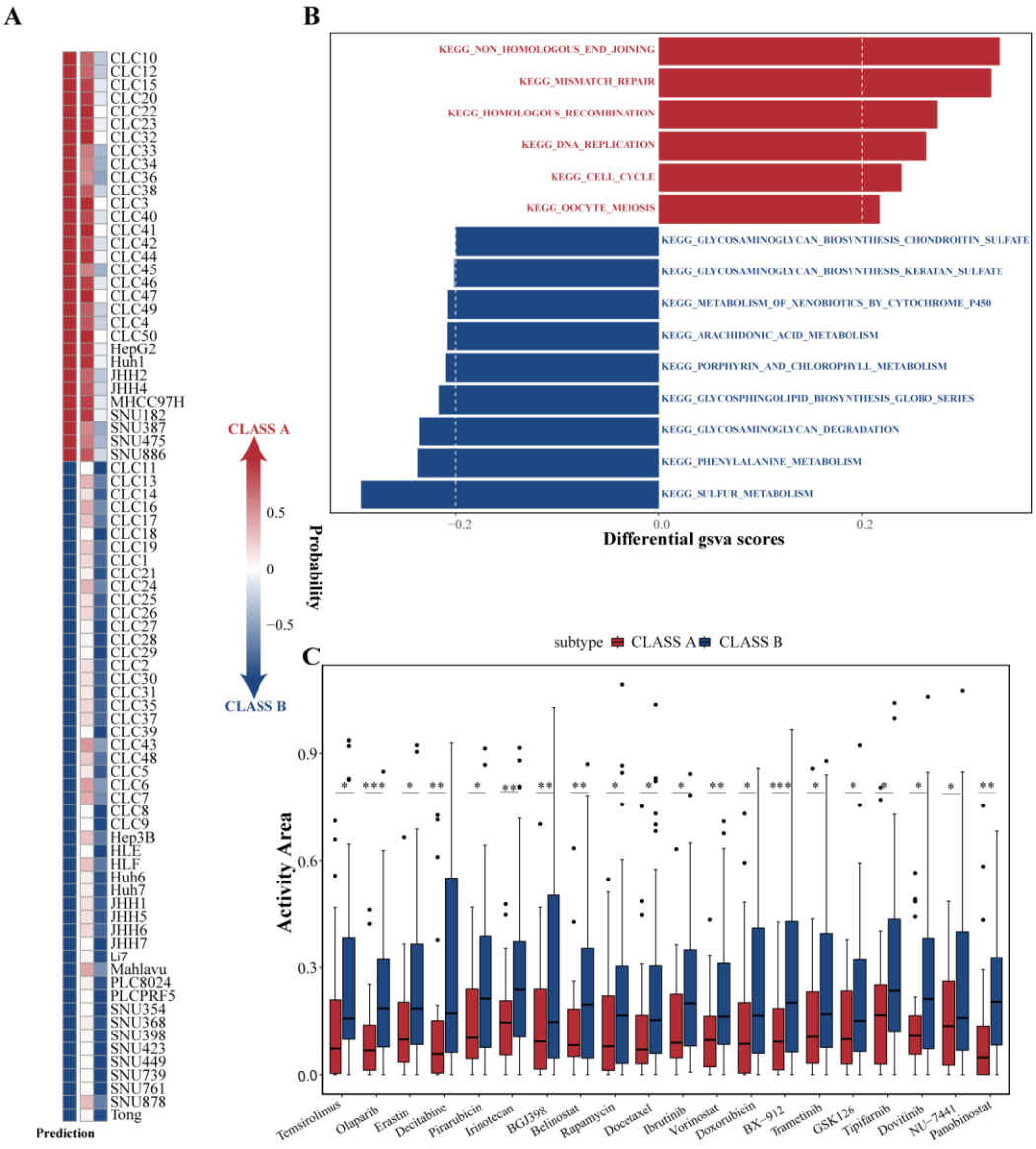


Figure 9 The drug sensitivity analysis of HCC subtypes. (A) The SVM_10 model assigned subtypes to 81 HCC cell lines from the LIMORE dataset. (B) KEGG enrichment analysis of the two subtypes. (C) Box plots illustrating drugs with differential activity area in the two subtypes.

## Reviewer #2

| 1.lack of description of the cohorts |
|---|

**Response**:

Thank you for reviewing our study. We have addressed your feedback about the lack of cohort description by implementing the necessary revisions. In the paper, we have included a new table(Supplementary Table S5) that extensively details sample information, clinical features, and other pertinent details for each cohort. This addition aims to provide readers with a more comprehensive understanding of the datasets we used, thereby bolstering the credibility and transparency of our study. We sincerely appreciate your valuable input, which has markedly contributed to enhancing the quality of our research.

Supplementary Table S5 Description of the patient cohort

| Clinical | Characteristics | TCGA_LIHC（n = 357） | ICGC_JP(n = 226) | ICGC_FR(n = 107) |
|---|---|---|---|---|
| Age | Median [1st Qu,3rd Qu] | 59 [52，68] | 67 [62，74] | 64 [57，73] |
| Gender | Male | 241 | 167 | 83 |
| | Female | 116 | 59 | 24 |
| Stage | I | 166 | 35 | NA |
| | II | 84 | 104 | NA |
| | III | 80 | 69 | NA |
| | IV | 5 | 18 | NA |
| | NA | 22 | 0 | NA |
| Grade | G1 | 51 | NA | NA |
| | G2 | 174 | NA | NA |
| | G3 | 116 | NA | NA |
| | G4 | 11 | NA | NA |
| | NA | 5 | NA | NA |
| Alcohol consumption | Yes | 113 | NA | NA |
| | No | 226 | NA | NA |
| | NA | 18 | NA | NA |
| HBV | Yes | 139 | NA | NA |
| | No | 218 | NA | NA |
| HCV | Yes | 99 | NA | NA |
| | No | 258 | NA | NA |
| Vascular invasion | Macro | 14 | NA | NA |
| | Micro | 89 | NA | NA |
| | None | 199 | NA | NA |
| | NA | 55 | NA | NA |
| AFP(ng/mL) | Median[1st Qu,3rd Qu] | 14251 [4, 282] | NA | NA |
| Platform | | Illumina HiSeq 2000 platform | Illumina HiSeq 2000 platform | Illumina HiSeq 2000 platform |
| Number of RNA-seq | | 60499 | 22913 | 57820 |

| Gene | | | |
|---|---|---|---|
| Country | USA | Japan | France |

2. no description of the selection for the 96 driver genes

**Response**:

We appreciate your question regarding the criteria for defining the initial 96 driver genes and the rationale for choosing these specific studies. We provide a supplementary explanation on this issue in Results Section 3.1. Here is our response:

First and foremost, we selected these three studies to obtain a more comprehensive and widely recognized list of HCC driver genes.

First, Bailey et al.'s[1] study, as part of the TCGA Pan-Cancer Atlas, employed a driver gene discovery algorithm based on single-gene mutation frequencies and a manual curation process to identify 299 driver genes. This study is widely regarded as a seminal work in the field of driver genes because it integrates data from multiple cancer types and provides a comprehensive gene list.

Second, the study by Martínez-Jiménez et al.[2], published in 2020, offers the most comprehensive catalogue of driver genes in cancer to date. While they did not involve domain experts in the curation of results, compared to Bailey et al.'s study, Martínez-Jiménez et al. analyzed a larger number of samples. They employed a more comprehensive approach to driver gene exploration, thus providing valuable insights.

Finally, Fujimoto et al.'s[3] study focused specifically on HCC, integrating the analysis of point mutations, structural variations, and viral integrations in both coding and non-coding regions of 300 HCC samples to identify HCC driver genes based on mutation frequencies. This study contributes to a better understanding of specific driver genes in HCC.

By selecting these three studies, we aimed to obtain a more comprehensive and widely recognized list of HCC driver genes to enhance our understanding of the driver gene mechanisms in HCC.

Reference:

[1] Bailey, Matthew H., et al. "Comprehensive characterization of cancer driver genes and mutations." Cell 173.2 (2018): 371-385.

[2] Martínez-Jiménez, Francisco, et al. "A compendium of mutational cancer driver genes." Nature Reviews Cancer 20.10 (2020): 555-572.

[3] Fujimoto, Akihiro, et al. "Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer." Nature genetics 48.5 (2016): 500-509.

3. lack of description of the RNAseq results

**Response**:

Thank you for your suggestions. In response, we have supplemented the RNA-seq results in the Methods and Results sections. For instance, we have included information on the initial gene count in the datasets and the retention of 17,660 genes after data preprocessing. Additionally, we have identified 2918 genes showing expression differences between normal and tumour tissues in our differential analysis. Furthermore, in the Results section, we have presented RNA-seq findings, specifically in Supplementary Table S3, where we describe the transcriptionally dysregulated genes

attributed to different driving factors such as DNA methylation, miRNA, and copy number variations.

In section 3.3 of the results, rather than employing traditional gene enrichment methods, we utilized the GSVA algorithm to evaluate the biological differences among different subtypes. GSVA utilizes expression information from all genes in the samples, not solely those differentially expressed. This approach allows a more comprehensive examination of biological pathway activities and aids in capturing even subtle expression differences among different samples. This is particularly crucial for understanding biological variations at the subtype or individual level."

This response provides a comprehensive explanation of the actions taken to address the issue raised by the reviewer regarding the description of the RNA-seq results.

| 4. No validation of the driver genes |
|---|

**Response**:

Thank you for your valuable suggestions. In fact, we have separately written a paper on the issues you mentioned, but unfortunately, it is still in the process of being published. Here are some preliminary results from our study on the extended list of driver genes:

After an in-depth exploration of the extended list of driver genes, we found that in humans, transcription factors constitute 8% of all genes and are associated with various diseases and phenotypes, with their mutations often carrying severe deleterious effects [1]. Previous studies have indicated frequent mutations in the zinc finger transcription factor family in endometrial cancer, colorectal cancer, and melanoma, contributing to widespread transcriptional dysregulation observed in cancer cells[2]. In our research, we observed an enrichment of mutations in the ZF-H2C2_2 domain in HCC, with nine mutation hotspots. The spatial structure of zf-H2C2_2 consists of an αββ structure formed by two consecutive C2H2-type zinc finger structures (Figure 1B). Through conservation analysis of mutation hotspots, we found that the majority of these hotspots are highly conserved (Figure 1A). Among them, the fifth and tenth residues of histidine are one of the four residues coordinating with zinc (Figure 1C), crucial for the stability of the domain, and substitutions in other residues may lead to the loss of domain function [3]. According to Barrera et al.'s study [4], we found that the residue at position 35 is connected to the DNA backbone, and its mutation may alter the binding affinity or specificity with DNA. Additionally, experimental studies have identified some zinc finger transcription factors closely associated with cancer progression. For example, ectopic expression of ZNF331 reduces the migration and invasion capabilities of cancer cells[5]. Mutations or translocations of ZNF687 are closely associated with the occurrence and development of cancer[6-7]. In HCC, the overexpression of ZNF687 significantly promotes liver cancer stem cell-like characteristics and tumor formation by upregulating the transcription levels of pluripotency-related factors BMI1, OCT4, and NANOG[8]. Mutations in the zinc finger domain of HNF4A often result in transcriptional dysregulation in HCC, promoting tumor growth[9]. The mutation frequency of these genes is extremely low (sometimes even less than 1%), making them challenging to identify in traditional single-gene driver gene identification algorithms. Martinez-Jimenez et al.'s study on a cohort of 1616 HCC samples could confidently identify only 1-2 members. The use of gene family extension can be considered a form of data augmentation, expanding the concept of repeated changes in individual genes in cancer to family domains with similar structures, thereby enhancing the statistical performance of the algorithm and uncovering more potential driver genes.

The main purpose of this manuscript is to classify HCC, explore the biological differences among different subtypes, and discover clinical diagnostic markers. We believe that including the validation of driver genes would make the entire manuscript excessively lengthy. Therefore, in the discussion section of this paper, we provided only a general description of this aspect. For example, we found that many significantly mutated domains are related to classical oncogenic signaling pathways, such as genes involved in the RTK signaling pathway, including the pkase_ty and SH2 domains. Downstream in the PI3K/AKT signaling pathway, domains such as PI3_PI4_kinase, PH, and PI3Ka are also enriched with mutations. Similarly, we observed the enrichment of mutations in the SET domain, a conserved catalytic core in the histone lysine methyltransferase family, crucial for the tumour-suppressive function according to [10].

Reference:


[1] Lambert, Samuel A., et al. "The human transcription factors." Cell 172.4 (2018): 650-665.

[2] Munro, Daniel, Dario Ghersi, and Mona Singh. "Two critical positions in zinc finger domains are heavily mutated in three human cancer types." PLoS computational biology 14.6 (2018): e1006290.

[3] Wolfe, Scot A., Lena Nekludova, and Carl O. Pabo. "DNA recognition by Cys2His2 zinc finger proteins." Annual review of biophysics and biomolecular structure 29.1 (2000): 183-212.

[4] Barrera, Luis A., et al. "Survey of variation in human transcription factors reveals prevalent DNA binding changes." Science 351.6280 (2016): 1450-1454.

[5] Yu, J., et al. "Zinc-finger protein 331, a novel putative tumor suppressor, suppresses growth and invasiveness of gastric cancer." Oncogene 32.3 (2013): 307-317.

[6] Divisato, Giuseppina, et al. "ZNF687 mutations in severe Paget disease of bone associated with giant cell tumor." The American Journal of Human Genetics 98.2 (2016): 275-286.

[7] Nguyen, TuDung T., et al. "Identification of novel Runx1 (AML1) translocation partner genes SH3D19, YTHDf2, and ZNF687 in acute myeloid leukemia." Genes, Chromosomes and Cancer 45.10 (2006): 918-932.

[8] Zhang, T., et al. "Overexpression of zinc finger protein 687 enhances tumorigenic capability and promotes recurrence of hepatocellular carcinoma." Oncogenesis 6.7 (2017): e363-e363.

[9] Taniguchi, Hiroaki, et al. "Loss-of-function mutations in Zn-finger DNA-binding domain of HNF4A cause aberrant transcriptional regulation in liver cancer." Oncotarget 9.40 (2018): 26144.

[10] Kim, Keun-Cheol, Liqing Geng, and Shi Huang. "Inactivation of a histone methyltransferase by mutations in human cancers." Cancer research 63.22 (2003): 7619-7623.


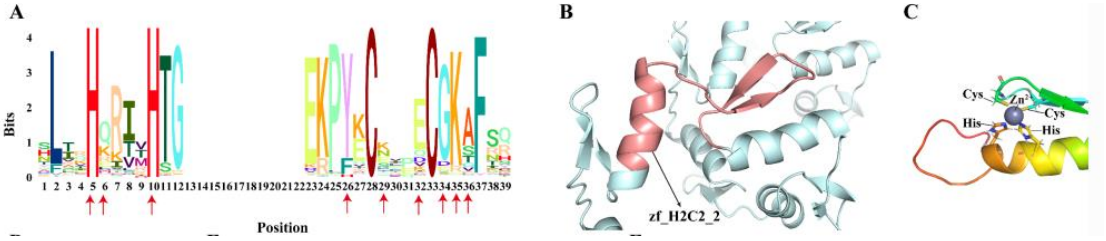


Fig. 1. Mutation hotspots in zf-H2C2_2 (A) Sequence logo of zf-H2C2_2 alignment(only shown position with significant mutations), red arrows indicate mutation hotspots (B) Spatial structure of zf-H2C2_2 (This structure is from Alpha Fold, and the low confidence region is removed) (C) Spatial structure of C2H2 zinc finger domains(PDB: 2eoz)

5. Finally, what is the rational of including microbiome data?

**Response**:

We appreciate your suggestion and in our discussion section, we have provided the rationale for analyzing microbiome data in our study. The polymorphic microbiome has been recognized for its association with prominent cancer hallmarks such as sustaining proliferative signalling, resisting cell death, and enabling replicative immortality[1]. Several studies have found a significant link between a polymorphic microbiome and cancer progression. For example, Jin et al.[2] identified that the local microbiota in lung cancer triggers inflammation associated with lung adenocarcinoma by activating γδ T cells within the lung. Mice treated with sterile conditions or antibiotics showed a marked reduction in lung cancer development caused by Kras mutation and p53 deletion. Similarly, Bullman et al.[3] revealed a close correlation between *Fusobacterium* and the progression of colorectal cancer, with antibiotic treatment reducing the *Fusobacterium* load and inhibiting tumour growth in colon cancer Patient-Derived Xenograft mice.

In HCC, the gut microbiota influences the liver through metabolites, subsequently impacting immune cell functionality[4]. In 2020, Poore et al.[5] assessed microbial abundance in TCGA cancer cohorts by analyzing genomic sequencing and RNA-seq raw data. They emphasized the distinct microbial compositions in each cancer type and substantial differences between cancers and normal samples. However, these differences were not significant within the same cancer type, such as between tumors at different stages. Given the unavailability of microbiome data in our study cohort, we leveraged Poore et al.'s results to explore differences in microbial characteristics among subtypes. While Poore et al.'s findings cannot replace the clinical gold standard for microbiome detection, our results suggest a trend – that microbial features significantly differ among subtypes of the same cancer. This trend may hold potential for diagnostic and therapeutic applications in cancer, though further exploration by experts in the field is warranted.

Reference:


[1] Hanahan, Douglas. "Hallmarks of cancer: new dimensions." Cancer discovery 12.1 (2022): 31-46.

[2] Jin C, Lagoudas G K, Zhao C, et al. Commensal microbiota promote lung cancer development via γδ T cells[J]. Cell, 2019, 176(5): 998-1013. e16.

[3] Bullman, Susan, et al. "Analysis of Fusobacterium persistence and antibiotic response in colorectal cancer." Science 358.6369 (2017): 1443-1448.

[4] Ma C, Han M, Heinrich B, et al. Gut microbiome–mediated bile acid metabolism regulates liver cancer via NKT cells[J]. Science, 2018, 360(6391): eaan5931.

[5] Poore, Gregory D., et al. "Microbiome analyses of blood and tissues suggest cancer diagnostic approach." Nature 579.7800 (2020): 567-574.