Hepatocellular carcinoma in Mongolia delineates unique molecular traits and a mutational signature associated with environmental agents

Laura Torrens[1,2#], Marc Puigvehí[1,3#], Miguel Torres-Martín[1,2], Huan Wang[4], Miho Maeda[1], Philipp K. Haber[1], Thais Leonel[5], Mireia García-López[5], Roger Esteban-Fabró[1,3], Wei Qiang Leow[1,6], Carla Montironi[1,2], Sara Torrecilla[2], Ajay Ramakrishnan Varadarajan[4], Patricia Taik[4], Genís Campreciós[1,5], Chinbold Enkhbold[7], Erdenebileg Taivanbaatar[8], Amankyeldi Yerbolat[8], Augusto Villanueva[1], Sofía Pérez-del-Pulgar[5], Swan Thung[1], Jigjidsuren Chinburen[8], Eric Letouzé[9], Jessica Zucman-Rossi[9], Andrew Uzilov[4,10], Jaclyn Neely[11], Xavier Forns[5], Sasan Roayaie[12], Daniela Sia[1], Josep M. Llovet[1,2,13*]

[1]Liver Cancer Program, Division of Liver Diseases, Tisch Cancer Institute, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, New York, USA

[2]Translational research in Hepatic Oncology, Liver Unit, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Hospital Clínic, University of Barcelona, Barcelona, Spain

[3]Hepatology Section, Gastroenterology Department, Parc de Salut Mar, IMIM (Hospital del Mar Medical Research Institute), Barcelona, Catalonia, Spain

[4]Sema4, Stamford, Connecticut, USA

[5]Liver Unit, Hospital Clínic, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), CIBEREHD, University of Barcelona, Barcelona, Spain

[6]Department of Anatomical Pathology, Singapore General Hospital, Singapore, Singapore

[7]Hepato-Pancreatico-Biliary Surgery Department, National Cancer Center, Ulaanbaatar, Mongolia

[8]National Cancer Center, Ulaanbaatar, Mongolia

[9]Centre de Recherche des Cordeliers, Sorbonne Université, Inserm, Université de Paris, Université Paris 13, Functional Genomics of Solid Tumors laboratory, F-75006, Paris, France

[10]Department of Genetics and Genomic Sciences and Icahn Institute for Data Science and Genomic Technology, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA

[11] Bristol Myers Squibb, Princeton, New Jersey, USA

[12]Department of Surgery, White Plains Hospital, White Plains, New York, USA

[13]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain

[#] These authors contributed equally

* Josep M. Llovet, MD, PhD

Mount Sinai Liver Cancer Program, Division of Liver Diseases, Icahn School of Medicine at Mount Sinai, 1425 Madison Avenue, Box 11-23, New York, NY 10029, USA

Tel.: +1 2126599503; fax: +1 212 849 2574

Email address: josep.llovet@mssm.edu

SUPPLEMENTARY MATERIALS AND METHODS

Clinical and histological data

All samples in the study were fresh-frozen. Tissue samples were coded previously to storage using consecutive numbering. The code did not include any patient identifier, and the research team at Mount Sinai received already de-identified samples. The diagnosis of HCC was confirmed after a first evaluation made by 3 independently working expert pathologists in Mount Sinai (WQL, CM, and ST), and those samples with >50% necrotic tissue (n=18), tumors other than HCC (n=7), or repeated (n=2) were excluded. Thus, a final number of 192 patients were included for further evaluation.

Baseline clinicopathological characteristics were collected for both cohorts. All histological evaluations were performed by 2 expert pathologists, blinded to clinical data. Fibrosis stage was scored according to the METAVIR Scale[1]. All the above-mentioned variables were also collected for the Western cohort except for BMI, alanine aminotransferase (ALT) values, tumor size and number, region of origin, and presence of steatosis/steatohepatitis. All data were stored in a database containing de-identified information, and electronic files were stored according to Mount Sinai IRB protocols with encryption and password protection.

Viral hepatitis evaluation

The presence of viral infections (HBV, HCV and HDV) was assessed in the non-tumor tissue of all Mongolian samples, and was compared to the data obtained from Mongolian charts (HBV surface antigen -HBsAg- and HCV antibodies; HDV was not routinely evaluated in Mongolia).

Intrahepatic HBV and HDV status were assessed by quantitative PCR (qPCR). HBV-DNA was assessed by Taqman qPCR (ID Pa03453406 s1, ABI, Thermo Fisher) using the ViiA7 Real Time PCR System (ABI) as previously described[2]. The calibration curve was prepared using ten-fold serial dilutions of a plasmid containing an HBV monomer (pHBV-EcoR1). Total HDV-RNA was determined by one-step RT-qPCR as previously reported[3,4]. For absolute quantification, serial dilution of an HDV-RNA standard (WHO 1st International Standard, Paul-Ehrlich-Institut) was included in each assay[5]. All samples positive for HDV were considered HBV/HDV positive.

HCV status was determined by conventional PCR. Specifically, HCV RNA was retrotranscribed to cDNA with EcoDry Premix (Double Primed) (Takara cat# 639549) and HCV-specific sequences were amplified under standard conditions using the following primer pair: Fw CACGCAGAAAGCGTCTAG, HCV; Rv TTGATCCAAGAAAGGACCC[6]. PCR products were run on an agarose gel, purified using PureLink Quick Gel Extraction Kit (Invitrogen cat# K210012) and sequenced by Sanger (Macrogen, USA).

HBV and HDV genotyping

HBV and HDV genotypes were determined by direct sequencing and phylogenetic analysis of a 1100 bp fragment of the HBV retrotranscriptase[7] and a fragment of 370bp encompassing approximately 85% of the large HDV antigen (HDAg)[8], respectively. Multiple alignments were performed with ClustalW[9] and maximum likelihood trees were obtained with MEGA X software[10].

Analysis of HBV mutations

HBV mutations associated with HCC development were assessed by nested PCR (GoTaq Flexi DNA Polymerase - Promega). Specifically, precore region was screened for nucleotide substitution G1896A and the basal core promoter (BCP) region was checked for the presence of 2 nucleotides substitutions (A1762T and G1764A). A DNA segment composing of the BCP, precore, and partial C regions was amplified by nested PCR and analyzed by direct sequencing[11].

Whole exome sequencing mutational variant calling in in-house cohorts

Mutational variant calling was performed following the Tigris pipeline (v2.0.1). BWA 0.7.17 was used for alignment, followed by base quality score recalibration via BQSR, read deduplication via Picard MarkDuplicates, germline molecular variant (SNV and small indel) calling via HaplotypeCaller, and somatic molecular variant calling via Mutect2, which calls variants using local de novo assembly and then does a two-pass filter using heuristics (further details can be found in the MuTect2 whitepaper from its GitHub repo at https://github.com/broadinstitute/gatk/tree/master/docs/mutect) . After applying these filters in MuTect2, the twice-filtered MuTect2 output was then filtered for 'PASS' variants only with allele frequency >= 5% for downstream analysis. Tigris computes depth-based and other NGS library QC metrics using GATK3 DepthOfCoverage and CallableLoci, as well as Picard. Lastly, somatic copy number variants (sCNV) were called using tumor/normal SAAS-CNV (v0.3.4) workflow that models allele balance to determine balanced versus unbalanced somatic gains and losses, as well as determine somatic copy-neutral loss of

heterozygosity[12]. SAAS-CNV output were further processed using GISTIC2.0 for somatic CNV analysis.

Analysis of gene mutations and filtering in previously published cohorts

We used whole exome sequenxing (WES data to assess the mutation profile in the European[13], Korean[14], TCGA[15] and Mongolian NCI[16] HCC cohorts. In the TCGA cohort, only variants with filter PASS were considered and "3_prime_UTR_variant", "5_prime_UTR_variant", "intron_variant"," synonymous_variant" were filtered from the cohort. For the Korean and European cohorts, only were accepted the following types of mutations, filtering the rest of the annotated subtypes: Missense_Mutation, "Nonsense_Mutation", "Splice_Site", "Translation_Start_Site", "Frame_Shift_Ins", "In_Frame_Ins", "Frame_Shift_Del", "In_Frame_Del", "3'Flank", "5'Flank" and "Nonstop_Mutation". For all cohorts, only VAF ≥ 0.05 was accepted for further analysis. Tumor mutational burden (TMB) from these external cohorts was calculated as previously indicated. Other TMB calculation approaches are provided in **Supplementary Table 7** for comparison.

Somatic copy number variations (SNVs) analysis

HaplotypeCaller[17] was used to generate germline VCF files as input for SAAS-CNV (v0.3.4)[12], which in turn generated segmentation file as input for GISTIC 2.0 run[18]. The "log2ratio.Median.adj" column from saasCNV output was used for GISTIC 2.0 run, with the following parameter flags -genegistic 1 -smallmem 1 -broad 1 -brlen 0.98 -conf 0.99 -armpeel 0 -savegene 1 -gcm extreme -qvt 0.1 -cap 2.0 -ta 0.85 -td 0.74.

Identification of potential driver genes

OncodriveCLUSTL and dN/dScv algorithms were used to identify genes harboring significantly more mutations than expected by chance[19,20] among the genes significantly more mutated in the Mongolian cohort compared to the Western cohort (n = 100, **Supplementary Table 10**). Genes predicted to have an enrichment for damaging alterations by OncodriveCLUSTL or dN/dScv were selected (q<0.05). The selected genes were filtered for cancer-related genes according to the OncoKB Cancer Gene List or previously reported studies in HCC[13,21].

TERT promoter mutations detection

The promoter region of *TERT* in Mongolian samples was amplified by PCR and sequenced using Sanger sequencing as previously described[22]. The number of *TERT* promoter mutations was compared to the reported percentages in Western cohorts (55-60%)[13].

Identification of de novo mutational signatures in Mongolian tumors

The MutationalPatterns[23] R package was used to perform *de novo* mutational signature extraction. Extracted signatures were mapped against COSMICv3. *De novo* signatures were mapped to single signatures and linear combinations of two if the cosine similarity was > 0.9. One novel signature "SBS Mongolia" was revealed with cosine similarity below the threshold for all comparisons (maximum observed cosine similarity of 0.818).

Mutational signature fitting was performed using the deconstructSigs (https://github.com/raerose01/deconstructSigs) and quadprog R packages[24], using HCC specific COSMICv3 mutational signatures plus SBS Mongolia. To select HCC specific

signatures, COSMICv3 signatures were assessed in 493 HCC samples from the Mongolian (n=151), Western (n=112) and TCGA (n=230) cohorts. Signatures occurring in ≥ 40 HCC samples or signatures that were revealed via *de novo* mutational signature extraction and able to be mapped to COSMICv3 reference were selected (i.e. SBS1, SBS4, SBS5, SBS6, SBS12SBS16, SBS18, SBS22, SBS26, SBS29, SBS40).

In order to assess the confidence of signature assignment across our samples in signature fitting, a previously reported bootstrap approach was adopted[25]. At each bootstrap, we randomly selected the same number of mutations with replacement from the original observed mutational profile of a given tumor sample (classified by the 96 trinucleotide mutation types) and performed signature fitting to estimate signature weight (quadprog R package), resulting in a distribution of signature weights for each signature from all bootstraps (N = 500) in a given tumor. Based on the signature weight distribution, for any given sample, we were able to estimate confidence level. At p value = 0.1 (one sided), the 10% quantile of signature weights would mean we were 90% confident that the signature weight was above that 10% quantile value. Finally, samples were considered positive for a mutational signature when the bootstrap exposure cutoff was ≥ 0.1.

Analysis of environmental signatures

Signature fitting analysis was performed using signatures from the Compendium of Mutational Signatures of Environmental Agents[26]. Specifically, all the 52 signatures included in the Compendium from agents generating significantly different substitution profiles compared to untreated controls were used[26]. The weights of each mutational signature contributing to an individual tumor sample were obtained using the deconstructSigs R

package. The trinucleotide count for each sample was normalized by multiplying it by a ratio of its occurence in the genome to its occurence in the exome (exome2genome method), following recommendations for WES data. Signature contributions with a weight <0.25 were discarded from the analysis. A signature was considered present in an individual tumor sample when the weight threshold was ≥ 0.1. (**Supplementary Table 22**).

Mutational signature analysis in external cohorts

Mutational signature fitting analysis was performed in the Korean[14] and Mongolian NCI[16] HCC cohorts as described above using COSMICv3 signatures plus SBS Mongolia, or environmental signatures from the Compendium of Mutational Signatures of Environmental Agents[26]. The Korean cohort was chosen as an example of Asian HCC cases likely not exposed to the same environmental factors as Mongolian patients.

Identification of de novo mutational signatures in Mongolian non-tumoral liver samples

Mutational signature analysis was used to assess signatures in the adjacent non-tumoral samples. First, for variant calling in the adjacent non-tumoral liver tissue, we subtracted the mutations in tumors from the mutation in non-tumoral tissue using MuTect2. Next, only samples with total SNV count ≥ 10 (for variants in exome region only at allelic frequency cutoff of 0.05) were selected for subsequent mutational signature analysis, resulting in a total of 78 samples (64 Mongolian cohort plus 14 Western cohort). Due to the small number of unique SNVs in adjacent non-tumoral samples, the analysis was performed on pooled variants from each cohort. The mutational signature fitting analysis was performed using all HCC specific COSMICv3 signatures plus SBS Mongolia, or environmental signatures from the Compendium of Mutational Signatures of Environmental Agents[26].

SUPPLEMENTARY REFERENCES

1.   Bedossa P, Poynard T. An algorithm for the grading of activity in chronic hepatitis C. The METAVIR Cooperative Study Group. Hepatology. 1996;24:289–93.

2.   Malmström S, Larsson SB, Hannoun C, Lindh M. Hepatitis B viral DNA decline at loss of HBeAg is mainly explained by reduced cccdna load - down-regulated transcription of PgRNA has limited impact. PLoS One. 2012;7.

3.   Ferns RB, Nastouli E, Garson JA. Quantitation of hepatitis delta virus using a single-step internally controlled real-time RT-qPCR and a full-length genomic RNA calibration standard. J Virol Methods. 2012;179:189–94.

4.   Giersch K, Homs M, Volz T, Helbig M, Allweiss L, Lohse AW, et al. Both interferon alpha and lambda can reduce all intrahepatic HDV infection markers in HBV/HDV infected humanized mice. Sci Rep. 2017;7.

5.   Le Gal F, Brichler S, Sahli R, Chevret S, Gordien E. First international external quality assessment for hepatitis delta virus RNA quantification in plasma. Hepatology. 2016;64:1483–94.

6.   Balart LA, Perrillo R, Roddenberry J, Regenstein F, Shim KS, Shieh YS, et al. Hepatitis C RNA in liver of chronic hepatitis C patients before and after interferon alfa treatment. Gastroenterology. 1993;104:1472–7.

7.   Tong Y, Liu B, Liu H, Zheng H, Gu J, Liu H, et al. New universal primers for genotyping and resistance detection of low HBV DNA levels. Med. 2016;95:e4618.

8.   Le Gal F, Brichler S, Drugan T, Alloui C, Roulot D, Pawlotsky J-M, et al. Genetic diversity and worldwide distribution of the deltavirus genus: A study of 2,152 clinical strains. Hepatology. 2017;66:1826–41.

9. Larkin MA, Blackshields G, Brown NP, Chenna R, Mcgettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007;23:2947–8.

10. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol. 2018;35:1547–9.

11. Chen CH, Changchien CS, Lee CM, Hung CH, Hu TH, Wang JH, et al. Combined mutations in pre-S/surface and core promoter/precore regions of hepatitis B virus increase the risk of hepatocellular carcinoma: A case-control study. J Infect Dis. 2008;198:1634–42.

12. Zhang Z, Hao K. SAAS-CNV: A Joint Segmentation Approach on Aggregated and Allele Specific Signals for the Identification of Somatic Copy Number Alterations with Next-Generation Sequencing Data. Wang E, editor. PLOS Comput Biol. 2015;11:e1004618.

13. Schulze K, Imbeaud S, Letouzé E, Alexandrov LB, Calderaro J, Rebouissou S, et al. Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. Nat Genet. 2015;47:505–11.

14. Ahn S-M, Jang SJ, Shim JH, Kim D, Hong S-M, Sung CO, et al. Genomic portrait of resectable hepatocellular carcinomas: implications of RB1 and FGF19 aberrations for patient stratification. Hepatology. 2014;60:1972–82.

15. Ally A, Balasundaram M, Carlsen R, Chuah E, Clarke A, Dhalla N, et al. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell. 2017;169:1327-1341.e23.

16. Candia J, Bayarsaikhan E, Tandon M, Budhu A, Forgues M, Tovuu LO, et al. The genomic landscape of Mongolian hepatocellular carcinoma. Nat Commun. 2020;11.

17. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, O. CM, Auwera GA Van der, et

al. Scaling accurate genetic variant discovery to tens of thousands of samples. Biorxiv 201178. 2018;

18. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011;12:R41.

19. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell. 2017;171:1029-1041.e21.

20. Arnedo-Pac C, Mularoni L, Muiños F, Gonzalez-Perez A, Lopez-Bigas N, Schwartz R. OncodriveCLUSTL: A sequence-based clustering method to identify cancer drivers. Bioinformatics. 2019;35:4788–90.

21. Chakravarty D, Gao J, Phillips S, Kundra R, Zhang H, Wang J, et al. OncoKB: A Precision Oncology Knowledge Base. JCO Precis Oncol. 2017;2017:1–16.

22. Nault JC, Mallet M, Pilati C, Calderaro J, Bioulac-Sage P, Laurent C, et al. High frequency of telomerase reverse-transcriptase promoter somatic mutations in hepatocellular carcinoma and preneoplastic lesions. Nat Commun. 2013;4:2218.

23. Blokzijl F, Janssen R, van Boxtel R, Cuppen E. MutationalPatterns: Comprehensive genome-wide analysis of mutational processes. Genome Med. 2018;10.

24. Lynch AG. Decomposition of mutational context signatures using quadratic programming methods. F1000Research. 2016;5:1253.

25. Huang X, Wojtowicz D, Przytycka TM. Detecting presence of mutational signatures in cancer with confidence. Bioinformatics. 2018;34:330–7.

26. Kucab JE, Zou X, Morganella S, Joel M, Nanda AS, Nagy E, et al. A Compendium of Mutational Signatures of Environmental Agents. Cell. 2019;177:821-836.e16.