Title

Somatic mutation landscape reveals differential variability of cell-of-origin for primary liver cancer

Authors and affiliations

Kyungsik Ha[1], Masashi Fujita[2], Rosa Karlić[3], Sungmin Yang[1], Yujin Hoshida[4], Paz Polak[5], Hidewaki Nakagawa[2], Hong-Gee Kim[1*] and Hwajin Lee[1*]

[1]Biomedical Knowledge Engineering Laboratory, Seoul National University, Seoul 08826, South Korea. [2]Laboratory for Genome Sequencing Analysis, RIKEN Center for Integrative Medical Sciences, Tokyo 108-8639, Japan. [3]Bioinformatics Group, Department of Molecular Biology, Division of Biology, Faculty of Science, University of Zagreb, Horvatovac 102a, 10000 Zagreb, Croatia. [4]Liver Tumor Translational Research Program, Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA. [5]Department of Oncological Sciences, Icahn School of Medicine at Mount Sinai, 1425 Madison Ave., NY 10029, USA.

[*]Correspondence: hgkim@snu.ac.kr; hwajin2k@gmail.com

Abstract

Background: Primary liver tissue cancers display consistent increase in global disease burden and mortality. Identification of cell-of-origins for primary liver cancers would be a necessity to expand options for designing relevant therapeutics and preventive medicine for these cancer types. Previous reports on cell-of-origin for primary liver cancers was mainly from animal studies, and integrative research utilizing human specimen data was poorly established.

Methods: We analyzed a whole-genome sequencing data set for a total of 363 tumor and progenitor tissues along with 423 normal tissue epigenomic marks to predict the cell-of-origin for primary liver cancer subtypes.

Results: Despite the mixed histological features, the predicted cell-of-origin for mixed hepatocellular carcinoma / intrahepatic cholangiocarcinoma were uniformly predicted as a hepatocytic origin. Individual sample-level prediction revealed differential level of cell-of-origin heterogeneity depending on the primary liver cancer types, with more heterogeneity observed in intrahepatic cholangiocarcinomas. Additional analyses on the whole genome sequencing data of hepatic progenitor cells suggest these progenitor cells might not a direct cell-of-origin for liver cancers.

Conclusions: These results provide novel insights on the heterogeneous nature and potential contributors of cell-of-origin predictions for primary liver cancers.

Keywords: Cell-of-origin, Primary liver cancer, Whole genome sequencing, Epigenomics, Cancer genomics

Background

Primary liver cancers (PLCs) is one of the major cancer types with increasing global disease burden over the years, reaching incidence rates and mortality over 900,000 per year (1, 2). This high morbidity and mortality of PLCs is due to the complex nature of the disease and lacking effective diagnostics and treatment besides multi-kinase inhibitors, thus strongly emphasizing the importance of relevant researches on early diagnosis and extensive drug development. In line with this, several endeavored researches were performed on identifying suitable diagnostic markers and targeted therapy-based treatments for PLCs, including the whole genome and exome-level profiling (3). So far, recent comprehensive efforts on investigating the genomics of PLCs revealed novel insights about the major mutation signatures, sub-classifications, and recurrent somatic mutations in coding regions (*TERT, TP53, CTNNB1, KRAS, IDH1/2,* etc.) and noncoding regions (*NEAT1* and *MALAT1*), which some of them are driver mutations and may associate with the clinical outcomes (4, 5). More investigations are underway to fully unveil the mechanisms and processes behind the progression of PLCs.

One of the complex, unanswered questions associated with the progression of PLCs is the possible cell-of-origins (COOs) corresponding to the various subtypes. PLC not only represents classical hepatocellular carcinoma (HCC) subtype, comprising of ~90% of PLCs, but also includes mixed hepatocellular and cholangiocarcinoma (Mixed) and intrahepatic cholangiocarcinoma (ICC), which are the two cancer subtypes displaying biliary phenotype with a different extent. COOs for these subtypes might depend on the location of a tumor within the liver and the differential clinical status associated with each tumor, represented by individual-level variability of cancer progression. So far, *in-vitro* and *in-vivo* experiments strictly at animal models proposed possible COOs for different subtypes of PLCs, including

hepatocytes for HCCs, Mixed and ICCs, cholangiocytes for Mixed and ICCs, and bipotential hepatic progenitor cells (HPCs) for HCCs and ICCs (6). None of these are yet confirmatory due to the potential biases accompanied by cell cultures and genetic manipulation-based lineage-tracing animal model systems and lack of human level studies, and both evidences which indicates either differentiated cells or HPCs as a predominant COO for PLCs are present. For example, COO for HCCs were either reported as solely hepatocytes (7) or hepatocytes plus differentiated benign lesions derived from HPCs (8). For the COO for ICCs, hepatocytes which undergo conversion into cholangiocytes (9) or the billiary epithelial cells (10) were pointed out as possible options depending on the usage of different transgenic models. In addition, recent reports also suggest the possibility of de-differentiation of hepatocytes (7) and cholangiocytes (11) after the liver injury as potential sources of progenitor cells and PLCs, which further enhances the complexity of cellular origin for the liver cancer progression. Efforts on extrapolating these COO-related complexities by utilizing actual human cancer tissue data itself were scarce with one article partly visiting at a preliminary level (12), but no studies were yet performed in a fully comprehensive, inter-cohort manner. Thus, uncovering the major COOs matching to each subtype of PLCs and examining the potential variance of COOs across the tumors from different individuals remain highly necessary for the better understanding of the cancer progression for PLCs along with the early-stage diagnosis and possibly the treatment selection.

Here, we performed a computational approach to dissect out the possible COOs matching to each cancer subtype within PLCs and to interrogate possible individual tumor-level heterogeneity in COOs. For this, we analyzed the whole genome sequencing data from 320 of PLCs (256 HCCs, 8 Mixed, and 56 ICCs), 12 of extrahepatic biliary tract cholangiocarcinoma (BTCAs) based on the assumption that these cancer type would display predominant

cholangiocytic COO, and 31 of HPCs and colon adult stem cells for assessing the possibility as a common COO for PLCs, along with 423 of chromatin features at the epigenome-level (see "Methods"). Our study not only confirmed the role of chromatin marks associated with possible COOs in shaping the mutation landscape of PLCs, but also uncovering the differential contribution of each COO in different subtypes of PLCs.

Methods

Data

For most analyses in this study, we used somatic mutation data of whole-genome sequencing (WGS) from the NCC-Japan liver cancer (LINC-JP), RIKEN-Japan liver cancer (LIRI-JP), and Singapore biliary tract cancer (BTCA-SG) projects after acquiring permission of ICGC (http://icgc.org). LINC-JP and LIRI-JP data consisted of a total of 282 samples with the exception of some cases which displayed multifocal or hypermutations, and these data were subgrouped according to the histological types (256 HCCs, 8 Mixed, and 18 ICCs). Data from BTCA-SG were all extrahepatic cholangiocarcinoma samples consisted of 12 samples without any particular subgroups.

The raw files of these datasets were analyzed along the standard GATK pipeline (https://www.broadinstitute.org/gatk/) and somatic mutations were called with the MuTect algorithm (http://archive.broadinstitute.org/cancer/cga/mutect) (13).

In addition to the data sets listed above, WGS-derived somatic mutation profile from additional 31 stem/progenitor samples (10 hepatic progenitor cells and 21 colon adult stem cells) and 38 ICCs from previous studies (5, 14) were utilized for the analysis related to hepatic progenitor cells (Fig. 3, Additional file 1: Figure S7) or as an independent cohort for predicting the COO of ICCs (Additional file 1: Figure S3) and assessing viral-infection associated COO predictions

for ICCs (Additional file 1: Figure S6a). Somatic variants of these samples were called from a different method that was designed in each study comparing to the datasets we analyzed.

A total of 423 epigenomic data for chromatin feature selections, correlation analyses and COO prediction analyses was obtained from ENCODE (15) and NIH Roadmap Epigenomics Mapping Consortium (16). NIH Roadmap epigenomics data can be accessed through the NCBI GSE18927 in Gene Expression Omnibus site (https://www.ncbi.nlm.nih.gov/geo/). In addition, chromatin data for liver tissues derived from hepatitis virus infected patients (donor HPC8 and HPC17) was obtained from IHEC (https://epigenomesportal.ca/ihec/download.html).

To estimate the regional mutation density and average signal of chromatin features, autosomes were divided into each 1-megabase region except sectors containing low quality unique mappable base pairs, centromeres, and telomeres.

We calculated the frequency of somatic mutations and ChIP-seq reads in each 1-megabase region to figure out the regional mutation density and histone modification profiles. The value of DNase I peaks and replication was also used to calculate DNase I hypersensitivity and Repli-seq profiles in each 1-megabase region. All these calculations were performed using BEDOPS (17).

Principal coordinate analysis

PCOA was employed to represent similarity/dissimilarity of mutation frequency landscapes among the samples. Each sample was represented in a two-dimensional space consisting of principal coordinates 1 and 2 using a dissimilarity matrix, which reflected Pearson correlation coefficient among the samples.

Feature selection based on random forest algorithm

Our feature selection analysis applied a modified version used in the previous study (18). Briefly, training set of each tree was organized and the mean squared error and the importance of each variable were evaluated using out-of-bag data. To determine the ranking of importance for each variable, the values of each variable were randomly permuted and examined to each tree. The initial importance value of variable m was estimated by subtracting the mean squared error between the untouched cases and the variable-m-permuted cases. Eventually, the ranking of each variable was determined by averaging importance values of variable m in the entire tree. We constructed a total of 1000 random forest trees to predict regional mutation density from a total of 423 chromatin features and employed greedy backward elimination to pick out the top 20 chromatin marks. This method sequentially removed the chromatin marker with the lowest rank at each step. These random forest models were repeated 1000 times each. Generally, in our feature selection analysis, the mutation density was calculated by combining the samples corresponding to each cancer type.

Prediction of cell-of-origin by grouping of chromatin features

To predict cell-of-origin (COO) for individual samples, chromatin marks were subgrouped based on the aggregate sample-level feature selection results. As a first step, we selected significant chromatin cell types above the cutoff score from the feature selection results using aggregated samples corresponding to each cancer type (Fig. 1a). Subsequently, we added relevant cell types and grouped the chromatin marks according to each selected cell type to evaluate the effect of cell-type specific chromatin on explaining variability of mutational landscapes among samples. For predicting the COO for HCCs, we simply utilized the

160 importance ranking among variables from 423 chromatin features due to the fact that liver
161 chromatin features were the only major type in the aggregated feature selection results for
162 HCCs.

163

164 **Signature analysis of mutational processes**

165 Nonnegative matrix factorization (NMF) algorithm was employed to investigate mutation
166 signatures as described in previous study (19). This methodology was utilized by factoring out
167 frequency matrix of 96-trinucleotide mutation contexts from HCC, Mixed, ICC, BTCA-SG
168 and HPC samples.

169

170 **Gene expression analysis**

171 RNA-Seq experiments of HCC samples were performed previously (4), and the data had been
172 deposited in the European Genome-phenome Archive. The reads were aligned onto the
173 reference human genome GRCh37 using TopHat v2.1.1. Raw read counts per gene were
174 computed using HTSeq with the GENCODE v19 annotation. Differential gene expression
175 between hepatocytic- and non- hepatocytic-origin HCCs was analyzed using limma-voom
176 v3.26.9 (20). Gene set enrichment analysis (GSEA) was performed using the GSEAPreranked
177 v5 module on the GenePattern server (https://genepattern.broadinstitute.org).

178

179

180

181

Results

Aggregate Sample-level Correlations Between Chromatin Marks and Somatic Mutations of PLCs

Based on the previous findings about the close associations between the chromatin feature levels and regional variations in somatic mutation frequencies of tumor (18) and a number of precancerous lesions (21), we first hypothesized that the whole-genome mutation landscape of hepatocytic PLC subtype (HCCs) would exhibit closer relationship with the liver tissue (surrogate tissue for the hepatocytes) chromatin marks, whereas the mutation landscape of partial or fully biliary PLC subtypes (Mixed and ICCs) and the BTCAs would likely to display higher correlations with the chromatin marks from tissues containing either cuboidal or columnar epithelium (kidney, stomach or intestines as representative surrogate tissues for the cholangiocytes), depending on the extent of biliary phenotype and anatomical location. To examine differential associations among the mutation landscape for different subtypes of PLCs and the chromatin feature levels from normal tissues, we first employed a random-forest based feature selection method to identify the chromatin features responsible for explaining the possible variances in regional somatic mutation frequencies. To conduct the analysis, we utilized the 1-megabase window somatic mutation frequency data for three subtypes of PLCs (HCCs, Mixed and ICCs) and BTCAs at an aggregated sample level along with the 1-megabase window chromatin feature counts. As hypothesized, liver tissue chromatin marks served as major features with significance for HCCs, and stomach tissue chromatin mark served as the first-rank feature for ICCs and BTCAs ($P < 2.2e-16$, Mann-Whitney U-test between the first- and second-rank features of each PLC subtype; Fig. 1a). Surprisingly, liver tissue chromatin marks were major features explaining the regional mutation variation of Mixed albeit containing the biliary phenotype, implicating the unexpected skewness of possible COO

206 towards to the hepatocytes for the particular subtype. The lower variance explained score for
207 Mixed and ICCs comparing to the HCCs were at least in part likely due to the lower number
208 of the samples and the total mutation load (Additional file 1: Figure S1a, b), indicating that the
209 correlation between the liver tissue chromatin feature levels and the somatic mutation
210 landscape of Mixed is similar to that of HCCs. In line with these result, spearman correlations
211 between the regional mutation frequency of HCCs or Mixed and liver H3K4me1 chromatin
212 mark level was the highest among different possible surrogate tissues, whereas stomach
213 H3K4me1 chromatin mark level showed the highest correlation with the regional mutation
214 frequency of BTCAs. (Additional file 1: Figure S2a, b). Spearman correlation values among
215 the regional mutation frequency of ICCs and H3K4me1 of different tissues were overall low
216 without any particular tissue type dependent differences, possibly due to both the lower
217 mutation load and the possible variability in COOs which have been previously reported (12).
218 Similar to the spearman correlation results, regional quintile-based mean mutation density data
219 of HCCs and Mixed were relatively highly associated with the liver tissue H3K4me1 level
220 comparing to the H3K4me1 level of stomach tissues, while mean mutation data of ICCs and
221 and BTCAs display higher association towards the stomach tissue H3K4me1, with ICCs as a
222 lesser extent (Fig. 1b). Collectively, these results demonstrate that the COO-associated
223 chromatin features could delineate the relationships with the mutation landscape of PLCs and
224 BTCAs.
225
226 **Individual Sample-level Cell-of-origin Predictions**
227 To further assess the differential mutation landscapes and possible COOs of PLCs and BTCAs
228 at the individual sample level, we conducted random forest algorithm-based COO analysis for
229 each sample. This individual sample-based COO analysis exhibited the dominance of

hepatocytic predicted COO for HCCs and Mixed, in contrast to the BTCAs which showed stomach tissues (a proxy tissue for extrahepatic cholangiocytes) possibly as a major COO (Fig. 2a). For ICCs, however, more heterogeneity of COO prediction was observed, and both hepatocytes and proxy tissues for cholangiocytes (kidney and stomach) were shown to be possible major COOs. This COO prediction pattern displayed consistency between different ICC cohorts (Additional file 1: Figure S3), thus emphasizing the heterogeneous nature of COO for ICCs. Our results not only replicated earlier findings on the COO of HCCs, ICCs and extrahepatic distal cholangiocarcinoma (DCCs) (12), but also additionally providing novel aspects about the complete predominance of hepatocytic predicted COO for Mixed tumors (8/8) and the implication of cuboidal cholangiocytes near the canal of hering (kidney tissue chromatin mark as a surrogate) could be another major COOs for ICCs besides the hepatocytes. In addition, 6 HCC samples showed non-hepatocytic predicted COO, thus inferring the possible distinctiveness for the COO of HCCs which might be linked to the differential tumor pathology. Overall, our results suggest the predominant COO for the HCCs and Mixed would most likely to be hepatocytes. Also, our evidences point to the possibility of cholangiocytes as a predominant COO for BTCAs, whereas the COOs of ICCs would vary by individual samples. These results implicate the importance of anatomical locations on the possible COOs of PLCs and BTCAs.

Alongside with these result, principle coordinate analysis (PCOA) result revealed that the PLCs with hepatocytic predicted COO tend to aggregate as a cluster with all of the samples displaying principle coordinate 1 value over 0 (Additional file 1: Figure S4). In terms of the PLC subtypes, HCCs and Mixed samples were all contained within a cluster except for the ones with non-hepatocytic predicted COOs, whereas the ICCs and BTCAs were more spread out (Fig. 2b), reflecting the distinct mutation landscape patterns.

To demonstrate whether HCCs with non-hepatocytic predicted COO have a unique phenotype compared to the hepatocyte-origin HCCs, we analyzed the gene expression profiles. Among the non-hepatocytic- and hepatocyte-origin predicted HCC samples, tumor RNA-Seq data were available for 6 and 189 samples, respectively (4). A comparison of gene expression levels between them showed that 124 genes were up-regulated and 21 genes were down-regulated in non-liver-origin HCCs (FDR < 0.05, absolute logFC > 0.647; Additional file 1: Table S1). Interestingly, the upregulated genes included an epithelial cell marker *EPCAM* and a cholangiocyte-specific marker *KRT19* (Fig. 2c). Clustering analysis confirmed that HCCs with non-hepatocytic predicted COO were enriched in a cluster that expressed more *EPCAM* and *KRT19* (Fig. 2d). Gene set enrichment analysis showed that molecular pathways associated with bile acid synthesis, xenobiotic degradation, and hepatocyte nuclear factor were down-regulated in HCCs with non-hepatocytic predicted COO (Additional file 1: Figure S5). This result indicates that the functional similarity to hepatocyte was lower in HCCs with non-hepatocytic predicted COO compared to hepatocyte-origin HCCs. Collectively, mRNA expression in non-hepatocyte-origin predicted HCCs partly resembled that of biliary epithelial cells. We also compared hepatocyte- and non-hepatocyte-origin predicted HCCs in terms of clinical features (including tumor stage and survival), but we did not see statistically significant difference in these features, implying that the COO assignments for HCCs might be independent from the clinical prognosis.

Previous publication described the association between hepatitis virus infection status and the liver COO assignments without any subgrouping of the virus types (12). As of further investigation, we tested whether there are any hepatitis virus-type dependent tendencies to particular COOs and the associated variance explained scores for the somatic mutation landscape of PLCs. Upon grouping the PLCs with the hepatitis B virus (HBV) and hepatitis C

278 virus (HCV) infection status, our analysis revealed that HCCs and Mixed samples were mostly
279 assigned to hepatocytic predicted COO regardless of the either hepatitis virus infection status.
280 In contrast, COO predictions on HCV-infected ICCs displayed predominance towards
281 hepatocytic predicted COO (n=5, binomial probability of 0.08, two tailed) and HBV infected
282 ICCs mostly displayed non-hepatocytic predicted COO assignments (n=9, binomial probability
283 of 0.04, two tailed) (Additional file 1: Figure S6a, c). Furthermore, spearman correlation values
284 between the regional mutation frequency of aggregated samples grouped by HBV or HCV
285 infection status and the normal liver tissue H3K4me1 chromatin mark level was higher for the
286 HCV-infected ICCs comparing to any other ICCs with different virus infection status, and this
287 result was fully replicated when using the H3K4me1 chromatin marks derived from HBV or
288 HCV-infected liver tissues, thus ensuring more relevancy (Additional file 1: Table S2). In line
289 with these results, variance explained scores for the ICCs calculated by using a total of 9 cell
290 or tissue types, we discovered that the chromatin features with the highest level of variance
291 explained scores were derived from different tissues depending on the hepatitis infection status
292 of ICCs (HBV = kidney tissue, HCV = liver tissue, NBNC = stomach tissue) (Additional file
293 1: Figure S6b). Albeit limited number of virus infected ICC samples, our results implicate a
294 potential skewness of COO depending on the virus infection status, and a separate cohort level
295 study with larger number of samples is strongly warranted. In addition, these results also
296 reflects the previous findings in differential infectivity of HBVs and HCVs for cholangiocytes
297 (22, 23).
298
299 **Hepatic Progenitor Cells as a Possible Cell-of-origin for PLCs**
300 HPCs, so called as oval cells, are a progenitor cell type located inside the Canal of Hering with
301 both hepatocytic and cholangiocytic differentiation capacity and suspected as a possible COO

for PLCs. To examine the possibility of HPCs as a possible COO for different subtypes of PLCs, we performed the random forest feature selection analysis using somatic mutation frequency data of HPCs (14) at an aggregate sample level along with the epigenome feature counts. Results from this analysis demonstrated that the mutation landscape of HPCs cannot be explained adequately by the chromatin landscape, with variance explained scores for the top rank chromatin feature and the total 423 features were either below 0 or 25% (Fig. 3a). In contrast, mutation frequency data of colon stem cells (14) (counterpart stem cell type) at an aggregate sample level were explained by pre-existing set of chromatin features with variance explained score over 40% for the H3K9me3 rectal mucosa mark and over 60% for the total 423 features. Post-adjustment of mutation load for colon stem cells at the level of HPCs still showed chromatin marks derived from the rectal mucosa tissue as a top rank feature with over 28% variance explained score, implicating that the differential mutation load might not be a contributing factor for the distinct feature selection analysis results. These results infer distinct mutation landscape between the HPCs and other PLCs, and thus points out the possibility that HPCs might not be a direct COO of PLCs.

Mutation signature analysis on the somatic mutation landscape of HPCs were previously performed, and identified a specific age-associated mutation signature displaying correlation with the replication timing and the average chromatin level of cell lines registered in the ENCODE project (14). Based on these findings, we conducted the mutation signature analysis on the HPCs along with the PLCs and BTCAs. As predicted, we successfully extracted a resembling signature (signature D) to the age-associated signature previously identified in the HPCs with similar relative proportion level, along with the other three mutation signatures (Additional file 1: Figure S7a, b). Next, we assessed whether the proportion of signature D is correlating with the COO assignment for PLCs. As demonstrated in Fig. 3b, relative

326 contribution level of signature D was significantly lower for non-hepatocytic predicted COO
327 assigned HCCs and ICCs comparing to the hepatocytic predicted COO assigned HCCs / ICCs
328 and all of the HPCs. In line with this, several evidences point out that the correlation between
329 relative proportion of the mutation signature and the COO assignment was specific and
330 consistent for signature D. One is that the proportion of other three signatures (A, B and C)
331 were not significantly associated with the COO assignments for ICCs ($P > 0.57$) and two
332 signatures (A, B) weren't showing any any signification associations with the COO
333 assignments for HCCs, too ($P > 0.24$). Also, mutation type pattern of HPCs were more
334 comparable to the ICCs and BTCAs rather than the HCCs and Mixed, in contrast to the findings
335 on the skewness of COO assignment depending on the signature D status. Furthermore, major
336 proportion of the non-hepatocytic predicted COO samples were located in the lower quartile
337 for the signature D proportions (Additional file 1: Figure S7d). Collectively, these results
338 provide a novel perspective in terms of the possible importance of age-associated mutation
339 signature level on the COO assignment, and thus reflecting again the distinct mutation
340 landscape between the hepatocytic and non-hepatocytic predicted COO samples.
341
342 **Discussion**
343 In this paper, we applied random-forest machine learning algorithm and other computational
344 analyses to whole genome sequencing data of PLCs and epigenomics data derived from normal
345 tissues to elucidate unique association patterns between the two features and identify possible
346 cell-of-origin distribution for PLCs at the subtype and individual tumor tissue level. Results
347 from these analyses would help to understand the complex and heterogeneous nature of cancer
348 cell-of-origin and the contribution of chromatin marks on differential regional somatic
349 mutation landscape during the progression for various subtypes of PLCs.

Several recent studies support the idea of chromatin marks serving as a crucial factor in shaping the mutation landscape for several types of tumors (18, 21, 24). Consistent with this idea, our results show that the chromatin marks can explain the mutation landscape of PLCs at the subtype level, displaying variance explained scores in the range of 56% (ICCs) to 87% (HCCs). Moreover, the top chromatin marks associated with the mutational landscape of 256 HCCs were mostly derived from the liver tissue and the top correlative chromatin marks for 12 of BTCAs were from the stomach tissue, which also directly matches to the previous results from the HCCs and DCCs (12). One thing to note is the lower level of variance explained scores for ICCs comparing to any other PLC subtypes. We speculate that the potential contributor to these differences in variance explained scores might be the lower mutation load and the higher level of heterogeneity in COOs at the individual tumor tissue-level since the COOs for individual ICC tissues were the most heterogeneous among all subtypes of PLCs, although following the cell and epithelial types with respect to the anatomical locations of ICCs.

The COOs for PLCs were highly debated for a number of years not only due to the discovery of several types of HPCs (25, 26), but also the facultative regeneration of hepatocytes and cholangiocytes which mainly occurs during the inflammation (7, 11). Our results suggest towards the differentiated cells rather than progenitor or stem cells as origins for PLCs based on the findings that the normal liver (representing hepatocytes), kidney and stomach (surrogate for the cholangiocytes) tissues can mostly explain the COO of PLCs, and the somatic mutation profile from the HPCs is not adequately explained (variance explained score < 24.04) by the normal tissue chromatin marks, albeit the significance of non-hepatocytic predicted COO assignments in regards to the age-associated mutation signature-specific manner. Although our chromatin feature selection analysis did not contain any liver progenitor/stem cell chromatin marks, poor correlation between the mutational landscape of HPCs and the liver or stomach

chromatin marks may infer the distinctiveness of chromatin landscape between the differentiated cells/tissues and the progenitor/stem cells. Although we cannot fully reject the possibility that the HPCs are still the very first COO of PLCs, our results at least suggest that the major somatic mutation accumulation would most likely to happen on differentiated cells, not at the progenitor/stem cell level. Future assessment on the relationship between the chromatin marks derived from the HPCs and the mutational landscape of PLCs and HPCs might be a separate confirmatory study, although the limitation on the number of progenitor/stem cells directly from human liver and its purity are major hurdles for ChIP-seq or any other epigenomics assays to be performed.

Conclusions

In summary, our results on the COO of PLCs discovered several novel and heterogenous nature of COO distributions in different subtypes. We believe that these results address the novel aspects of individual-level differences in tumor biology and clinical pathology of PLCs, and providing a robust and relevant way of studying cancer COO in human system without utilizing a human organoid system, which might be solely suitable for mechanism studies in a practical manner due to the labor intensiveness caused by making each organoid per each patient and potential selection bias during cell culture. Ultimately, our results might add significant arguments for the necessity of personalized medicine for cancer treatments, combined with the genomics and the other molecular signatures.

Abbreviations

PLC: primary liver cancer; COO: cell-of-origin; HCC: hepatocellular carcinoma; Mixed: mixed hepatocellular and cholangiocarcinoma; ICC: intrahepatic cholangiocarcinoma; HPC: hepatic progenitor cell; BTCA: extrahepatic biliary tract cholangiocarcinoma; WGS: whole-genome sequencing; LINC-JP: NCC-Japan liver cancer; LIRI-JP: RIKEN-Japan liver cancer; BTCA-SG: Singapore biliary tract cancer; NMF: nonnegative matrix factorization; GSEA : gene set enrichment analysis; DCC: extrahepatic distal cholangiocarcinoma; PCOA: principle coordinate analysis; HBV: hepatitis B virus; HCV: hepatitis C virus

Acknowledgements

Not applicable

Funding

This work was partly supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea Ministry of Science, ICT and future Planning (MSIP) (No.2017-0-00398, Development of drug discovery software based on big data) and the National Research Foundation of Korea(NRF) funded by the MSIP (No. NRF-2017R1A4A1014584, Epigenetic Regulation of Bone & Muscle Regeneration Lab).

Availability of data and materials

WGS and RNA-seq data are deposited in the European Genome-phenome Archive, and these data can be accessed with the approval of the ICGC Data Access Compliance Office. NIH Roadmap epigenomics data can be obtained at Gene Expression Omnibus site (https://www.ncbi.nlm.nih.gov/geo/) under accession number NCBI GSE18927. In addition,

421 chromatin data of liver tissues derived from hepatitis virus infected patients (donor HPC8 and
422 HPC17) can be accessed at IHEC (https://epigenomesportal.ca/ihec/download.html).
423
424 **Authors' contributions**
425 HL provided the original idea. HL, KH, and HK led the overall project. KH, MF, RK, SY, and
426 HL analyzed the data and contributed to scientific discussions. HL and KH wrote the
427 manuscript, and MF, RK, PP, YH, HN, and HK reviewed the manuscript.
428
429 **Ethics approval and consent to participate**
430 Not applicable
431
432 **Consent for publication**
433 Not applicable
434
435 **Competing interests**
436 H.L. is currently working at UPPThera, Inc., but conducted the current research without any
437 conflict of financial interests. Other authors declare no competing financial interests.
438

References

1. Disease GBD, Injury I, Prevalence C. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet. 2016;388(10053):1545-602.

2. Mortality GBD, Causes of Death C. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet. 2016;388(10053):1459-544.

3. Ziogas DE, Kyrochristos ID, Glantzounis GK, Christodoulou D, Felekouras E, Roukos DH. Primary liver cancer genome sequencing: translational implications and challenges. Expert Rev Gastroenterol Hepatol. 2017;11(10):875-83.

4. Fujimoto A, Furuta M, Totoki Y, Tsunoda T, Kato M, Shiraishi Y, et al. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. Nat Genet. 2016;48(5):500-9.

5. Jusakul A, Cutcutache I, Yong CH, Lim JQ, Huang MN, Padmanabhan N, et al. Whole-Genome and Epigenomic Landscapes of Etiologically Distinct Subtypes of Cholangiocarcinoma. Cancer Discov. 2017;7(10):1116-35.

6. Sia D, Villanueva A, Friedman SL, Llovet JM. Liver Cancer Cell of Origin, Molecular Class, and Effects on Patient Prognosis. Gastroenterology. 2017;152(4):745-61.

7. Mu X, Espanol-Suner R, Mederacke I, Affo S, Manco R, Sempoux C, et al. Hepatocellular carcinoma originates from hepatocytes and not from the progenitor/biliary compartment. J Clin Invest. 2015;125(10):3891-903.

8. Tummala KS, Brandt M, Teijeiro A, Grana O, Schwabe RF, Perna C, et al. Hepatocellular Carcinomas Originate Predominantly from Hepatocytes and Benign Lesions

463 from Hepatic Progenitor Cells. Cell Rep. 2017;19(3):584-600.

464 9. Sekiya S, Suzuki A. Intrahepatic cholangiocarcinoma can arise from Notch-mediated
465 conversion of hepatocytes. J Clin Invest. 2012;122(11):3914-8.

466 10. Guest RV, Boulter L, Kendall TJ, Minnis-Lyons SE, Walker R, Wigmore SJ, et al. Cell
467 lineage tracing reveals a biliary origin of intrahepatic cholangiocarcinoma. Cancer Res.
468 2014;74(4):1005-10.

469 11. Raven A, Lu WY, Man TY, Ferreira-Gonzalez S, O'Duibhir E, Dwyer BJ, et al.
470 Cholangiocytes act as facultative liver stem cells during impaired hepatocyte regeneration.
471 Nature. 2017;547(7663):350-4.

472 12. Wardell CP, Fujita M, Yamada T, Simbolo M, Fassan M, Karlic R, et al. Genomic
473 characterization of biliary tract cancers identifies driver genes and predisposing mutations. J
474 Hepatol. 2018;68(5):959-69.

475 13. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al.
476 Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples.
477 Nat Biotechnol. 2013;31(3):213-9.

478 14. Blokzijl F, de Ligt J, Jager M, Sasselli V, Roerink S, Sasaki N, et al. Tissue-specific
479 mutation accumulation in human adult stem cells during life. Nature. 2016;538(7624):260-4.

480 15. Consortium EP. An integrated encyclopedia of DNA elements in the human genome.
481 Nature. 2012;489(7414):57-74.

482 16. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al.
483 Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317-30.

484 17. Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, et al.
485 BEDOPS: high-performance genomic feature operations. Bioinformatics. 2012;28(14):1919-
486 20.

487 18. Polak P, Karlic R, Koren A, Thurman R, Sandstrom R, Lawrence M, et al. Cell-of-
488 origin chromatin organization shapes the mutational landscape of cancer. Nature.
489 2015;518(7539):360-4.
490 19. Blokzijl F, Janssen R, van Boxtel R, Cuppen E. MutationalPatterns: comprehensive
491 genome-wide analysis of mutational processes. Genome Med. 2018;10(1):33.
492 20. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential
493 expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res.
494 2015;43(7):e47.
495 21. Ha K, Kim HG, Lee H. Chromatin marks shape mutation landscape at early stage of
496 cancer progression. NPJ Genom Med. 2017;2:9.
497 22. Fletcher NF, Humphreys E, Jennings E, Osburn W, Lissauer S, Wilson GK, et al.
498 Hepatitis C virus infection of cholangiocarcinoma cell lines. J Gen Virol. 2015;96(Pt 6):1380-
499 8.
500 23. Blum HE, Stowring L, Figus A, Montgomery CK, Haase AT, Vyas GN. Detection of
501 hepatitis B virus DNA in hepatocytes, bile duct epithelium, and vascular elements by in situ
502 hybridization. Proc Natl Acad Sci U S A. 1983;80(21):6685-8.
503 24. Polak P, Lawrence MS, Haugen E, Stoletzki N, Stojanov P, Thurman RE, et al.
504 Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair.
505 Nat Biotechnol. 2014;32(1):71-5.
506 25. Cardinale V, Wang Y, Carpino G, Cui CB, Gatto M, Rossi M, et al. Multipotent
507 stem/progenitor cells in human biliary tree give rise to hepatocytes, cholangiocytes, and
508 pancreatic islets. Hepatology. 2011;54(6):2159-72.
509 26. Wang B, Zhao L, Fish M, Logan CY, Nusse R. Self-renewing diploid Axin2(+) cells
510 fuel homeostatic renewal of the liver. Nature. 2015;524(7564):180-5.

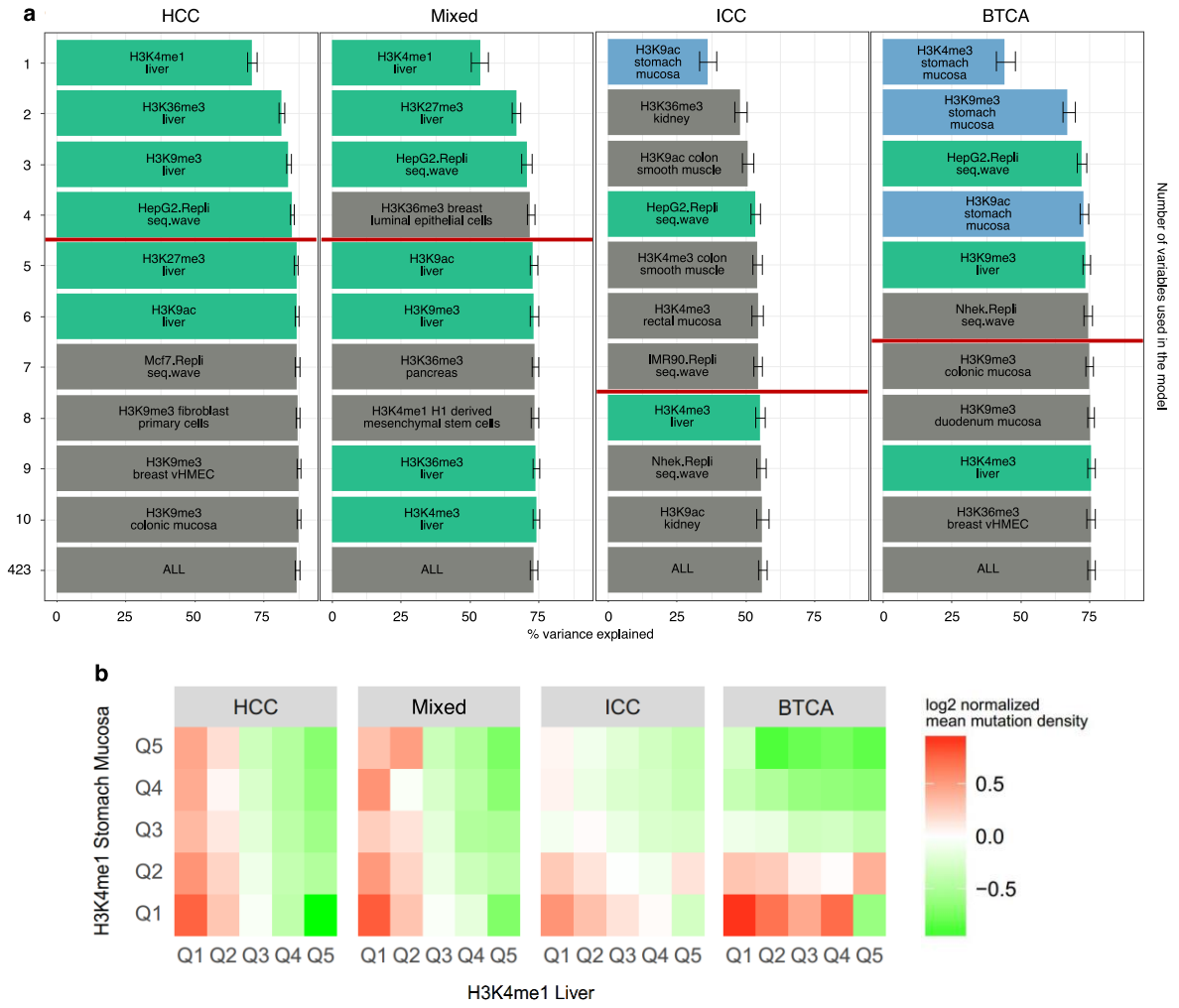

Fig. 1 Cell-of-origin chromatin features delineating relations with the regional mutation frequency of HCCs, Mixed, ICCs and BTCAs. a Random forest regression-based chromatin feature selection using aggregated somatic mutation frequency data from HCC, Mixed, ICC and BTCA-SG samples. The rank of each chromatin feature is determined by importance values. The bar length represents the variance explained scores, and error bar shows minimum and maximum scores derived from 1,000 repeated simulations. Red lines represent the cutoff scores determined by the prediction accuracy of 423 features-1 s.e.m. Liver chromatin features are green-colored and stomach chromatin features are blue-colored. **b** Normalized mean mutation density per each PLC subtype and BTCAs plotted with respect to the density quintile groups of liver and stomach H3K4me1 marks.

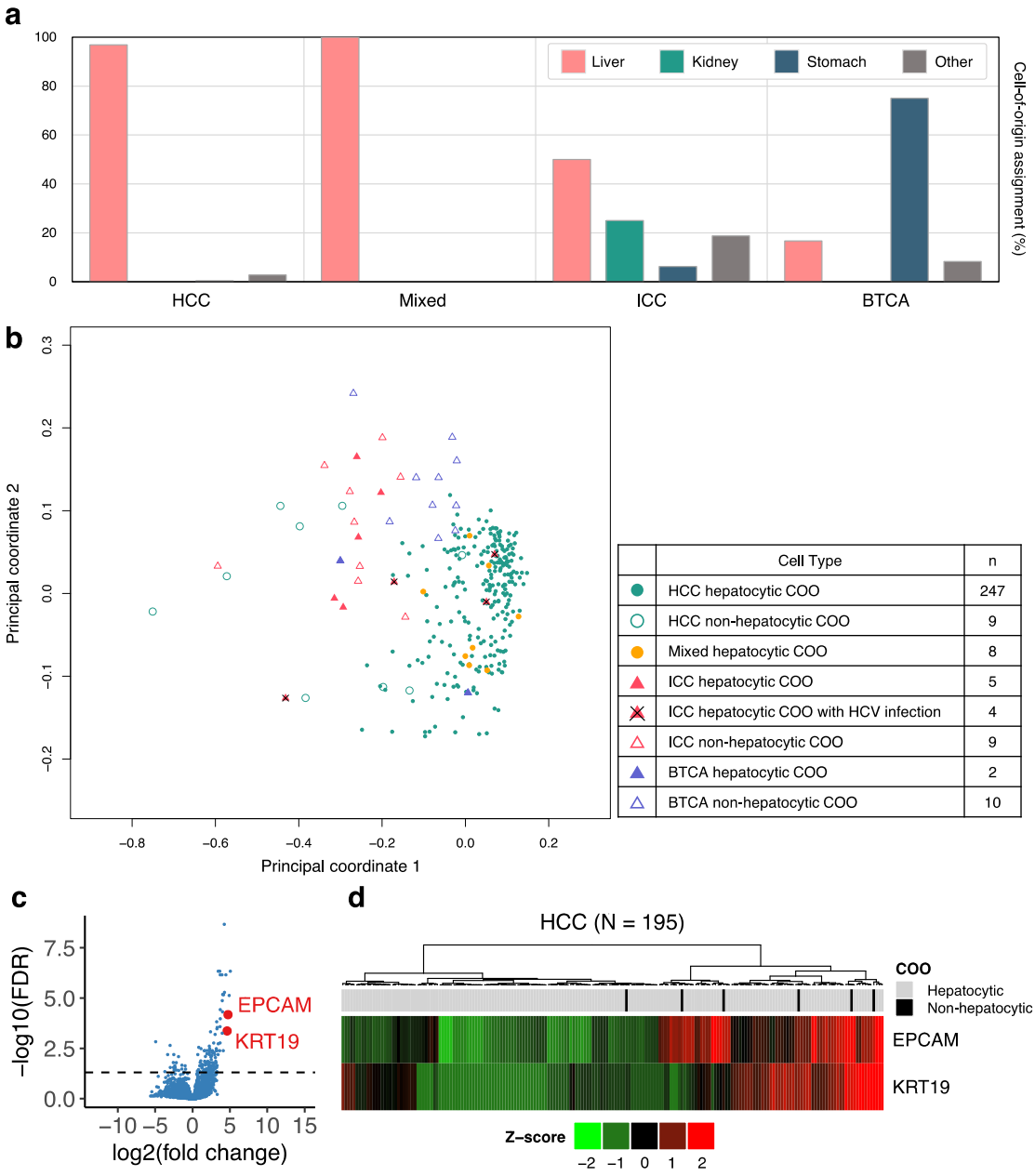

Fig. 2 Analysis of COOs at the individual cancer samples. a Prediction of COO via grouping of chromatin features for each normal tissue type. Bar graph represents the percentage of samples with respect to the assigned COO by liver tissue chromatin features (pink), kidney tissue chromatin features (green), stomach tissue chromatin features (navy) or the rest (gray). **b** Principal coordinate analysis (PCOA) of mutation frequency distributions for individual cancer samples. **c, d** Differential gene expression by non-hepatocytic COO HCCs (n = 6) comparing to the hepatocytic COO HCCs (n = 189). **c** Volcano plot. The horizontal axis is the log-ratio of the non-hepatocytic COO to the hepatocytic origins. Dashed line represents FDR = 0.05. **d** Expression profile of *EPCAM* and *KRT19* mRNA.

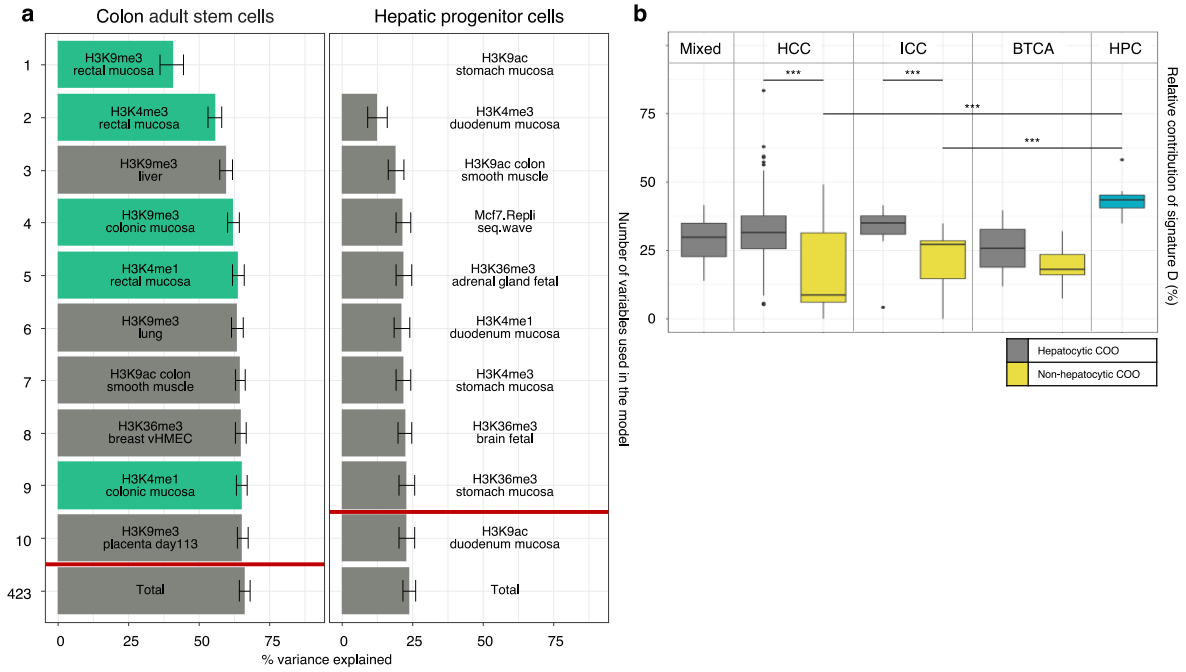

Fig. 3 Hepatic progenitor cells have distinct mutation landscape and mutational signature processes compared to primary liver cancer genomes. a Chromatin feature selection in relation to the regional mutation frequency of colon adult stem cells and hepatic progenitor cells. Chromatin features related to each tissue type are green-colored. **b** Box plot represents the distribution of relative contribution of signature D in HCC, Mixed, ICC, BTCA and HPC samples. Samples of each tumor type are separated based on whether they are predicted as hepatocytic COO (gray) or not (yellow). Statistical significance is calculated by using Mann-Whitney U-test (***, $P < 0.05$). BTCAs were excluded from the statistical analysis because only two samples were predicted as hepatocytic COO.

Additional file

Supplementary tables and figures for Ha et al. “Somatic mutation landscape reveals differential variability of cell-of-origin for primary liver cancer”.

Contents: Tables S1-S2 and Figures S1-S7

Table S1. Differentially expressed Genes between non-hepatocytic- and hepatocytic-origin HCCs.

Gene ID	Gene symbol	logFC	AveExpr	P.Value	FDR
ENSG00000161249.16	DMKN	5.062718005	-0.133833627	5.79731E-11	4.60909E-07
ENSG00000134121.5	CHL1	4.903538873	-2.745241348	4.5308E-09	7.38808E-06
ENSG00000119888.6	EPCAM	4.750318163	0.451955596	5.56417E-08	6.65364E-05
ENSG00000171345.9	KRT19	4.629835737	-2.242115976	5.04115E-07	0.000430586
ENSG00000162949.12	CAPN13	4.460917837	-3.452568356	2.51523E-10	6.95756E-07
ENSG00000219438.4	FAM19A5	4.276950727	-1.839672486	2.31201E-09	5.18381E-06
ENSG00000131037.10	EPS8L1	4.249681151	-0.736121173	1.19391E-13	2.14152E-09
ENSG00000184363.5	PKP3	4.154360112	-3.154846101	3.80302E-09	6.82148E-06
ENSG00000104413.11	ESRP1	4.112951635	-3.329175306	9.62504E-06	0.00411058
ENSG00000183454.9	GRIN2A	4.036506636	-3.619026899	3.32308E-09	6.62289E-06
ENSG00000159263.11	SIM2	4.00916387	-3.108501219	3.7529E-08	4.80827E-05
ENSG00000182272.7	B4GALNT4	3.980777347	-2.998102938	8.9587E-09	1.3391E-05
ENSG00000162069.10	CCDC64B	3.97193709	-3.223893094	7.1183E-07	0.000555135
ENSG00000146555.14	SDK1	3.941227041	-1.162653665	2.71522E-10	6.95756E-07
ENSG00000153404.9	PLEKHG4B	3.712740228	-2.930286008	9.4771E-08	0.000106244
ENSG00000162552.10	WNT4	3.704973809	-0.601769663	2.71779E-08	3.74993E-05
ENSG00000136002.12	ARHGEF4	3.704371314	-2.720236555	2.01164E-10	6.95756E-07
ENSG00000165238.12	WNK2	3.688090343	0.029657067	2.45255E-06	0.001691976
ENSG00000145113.17	MUC4	3.685707427	-2.621147752	2.63751E-07	0.000262828
ENSG00000137203.6	TFAP2A	3.647603437	-0.592925872	9.60929E-11	4.60909E-07
ENSG00000165449.7	SLC16A9	3.633684789	-0.41092064	8.417E-06	0.00411058
ENSG00000105048.12	TNNT1	3.44692295	-3.497199786	2.90823E-06	0.001863034
ENSG00000189292.11	FAM150B	3.440147878	-2.392342705	1.02784E-10	4.60909E-07
ENSG00000159247.8	TUBBP5	3.419449712	-2.84541423	1.91239E-05	0.006417903
ENSG00000111344.7	RASAL1	3.305183228	-2.063788982	3.03164E-07	0.000286203
ENSG00000184292.5	TACSTD2	3.284731657	-1.487481534	0.00010527	0.022478897
ENSG00000170074.15	FAM153A	3.271527857	-2.568659534	3.75069E-07	0.000336381
ENSG00000124102.4	PI3	3.266949152	-2.131724945	0.000344921	0.045176245
ENSG00000268756.1	AC104534.2	3.266112552	-3.219729752	9.8852E-06	0.004123509
ENSG00000133477.12	FAM83F	3.240905512	-2.952343337	0.000152303	0.02787617
ENSG00000184343.6	SRPK3	3.237933881	-3.06315791	1.42162E-07	0.000149997
ENSG00000112812.11	PRSS16	3.218769296	-3.298671945	0.000111322	0.023184212
ENSG00000104892.12	KLC3	3.192297988	-2.986348724	8.53829E-06	0.00411058
ENSG00000225946.1	RP11-395B7.2	3.179456656	-3.127704284	6.19947E-06	0.003270587
ENSG00000095932.5	C19orf77	3.153529438	-0.045867821	6.60897E-05	0.016239054
ENSG00000005001.5	PRSS22	3.101113529	-3.384871231	4.94786E-05	0.013867142
ENSG00000149043.12	SYT8	3.063792717	-2.314471678	2.98231E-05	0.009066726
ENSG00000069188.12	SDK2	3.025609327	-1.962029986	8.5414E-07	0.000638363
ENSG00000176920.10	FUT2	3.015073918	-1.667011629	0.000197405	0.032484932
ENSG00000171462.10	DLK2	2.992424532	-1.858245628	7.22012E-05	0.017500984
ENSG00000166796.7	LDHC	2.981548718	-2.831651639	1.80405E-05	0.006417903
ENSG00000187775.12	DNAH17	2.97321331	-1.386054121	0.000164151	0.028105689
ENSG00000130294.10	KIF1A	2.942405647	-1.873754387	0.000164526	0.028105689
ENSG00000135373.8	EHF	2.893291719	1.242929185	1.3366E-05	0.005181954
ENSG00000013588.5	GPRC5A	2.88411355	-1.717433767	0.000162077	0.028105689
ENSG00000188112.4	C6orf132	2.859416155	-1.931629437	0.000202608	0.033037931
ENSG00000105426.10	PTPRS	2.836523141	1.309481792	2.57035E-06	0.001707569
ENSG00000117322.12	CR2	2.834777009	-3.013454749	0.000238513	0.035933308
ENSG00000137699.12	TRIM29	2.803176897	-0.712695	9.93815E-05	0.021739101

Gene ID	Gene symbol	logFC	AveExpr	P.Value	FDR
ENSG00000154319.10	FAM167A	2.789097694	-2.048367329	1.21079E-05	0.004935904
ENSG00000159212.8	CLIC6	2.739799056	-2.06770182	0.000323615	0.043644245
ENSG00000205795.4	CYS1	2.715764216	-2.602504102	0.000114339	0.023184212
ENSG00000101115.8	SALL4	2.704137018	0.268588921	0.000261316	0.037800136
ENSG00000058404.15	CAMK2B	2.683984591	-0.902376697	6.39936E-05	0.016179566
ENSG00000102554.9	KLF5	2.676343614	1.740264907	1.35782E-05	0.005181954
ENSG00000185499.12	MUC1	2.661448334	0.281755422	9.22351E-06	0.00411058
ENSG00000204380.2	AC005042.4	2.658207752	-3.35332207	0.0003644	0.046356287
ENSG00000198753.7	PLXNB3	2.580845097	-0.342017577	2.22385E-05	0.007252568
ENSG00000162738.5	VANGL2	2.572483777	-0.947963707	0.000224222	0.034083666
ENSG00000131203.8	IDO1	2.571589247	0.913726831	0.000123594	0.023335931
ENSG00000170425.3	ADORA2B	2.524458011	-1.078393131	0.000112772	0.023184212
ENSG00000181218.4	HIST3H2A	2.510548398	-2.210642505	0.000122127	0.023304256
ENSG00000182580.2	EPHB3	2.509108717	-0.632213464	3.81808E-05	0.011414155
ENSG00000168453.10	HR	2.470199317	-1.559337528	0.000141445	0.026428178
ENSG00000196155.8	PLEKHG4	2.456633551	0.21091219	2.7527E-05	0.00851295
ENSG00000101213.5	PTK6	2.428244737	0.647886993	0.000170132	0.028789256
ENSG00000143797.7	MBOAT2	2.410345408	0.613264802	5.78745E-07	0.000471861
ENSG00000167642.8	SPINT2	2.37788438	2.136207673	0.000209611	0.03353451
ENSG00000181085.10	MAPK15	2.324487437	-2.43292748	0.000355599	0.045887569
ENSG00000205363.4	C15orf59	2.316898483	-2.449986115	0.00036084	0.046231286
ENSG00000130751.5	NPAS1	2.302061395	-1.489840859	8.7091E-06	0.00411058
ENSG00000181652.14	ATG9B	2.299726617	-1.625076203	6.13289E-06	0.003270587
ENSG00000112655.11	PTK7	2.295893369	1.858078697	0.000120206	0.023184212
ENSG00000249684.1	RP11-423H2.3	2.289066134	-1.995452163	5.92996E-05	0.015821634
ENSG00000143320.4	CRABP2	2.286963455	-0.737104636	6.56307E-05	0.016239054
ENSG00000228594.1	C1orf233	2.260768638	-0.826738845	1.7885E-05	0.006417903
ENSG00000126460.6	PRRG2	2.259200701	0.458296183	1.32821E-05	0.005181954
ENSG00000155066.11	PROM2	2.251707637	-1.514406248	0.000117128	0.023184212
ENSG00000072071.12	LPHN1	2.248413263	1.363492015	4.10616E-06	0.002455075
ENSG00000171219.8	CDC42BPG	2.244840337	0.912398749	0.000103225	0.02230782
ENSG00000168350.6	DEGS2	2.225973983	-1.469016498	0.000368803	0.046586032
ENSG00000169583.12	CLIC3	2.2189328	-1.793956755	0.000153872	0.027878816
ENSG00000007171.12	NOS2	2.156946925	-0.368374591	8.71634E-05	0.020044234
ENSG00000143882.5	ATP6V1C2	2.135760717	-0.275467443	0.00028698	0.039903597
ENSG00000145287.6	PLAC8	2.128659333	0.242485682	8.51542E-05	0.019836511
ENSG00000167608.7	TMC4	2.126594095	1.449640447	0.000180133	0.029917043
ENSG00000124466.8	LYPD3	2.093261972	-0.6772654	0.000284823	0.039903597
ENSG00000164114.14	MAP9	2.093124378	-0.682488736	0.000211262	0.03353451
ENSG00000180787.5	ZFP3	2.05134027	0.564148684	4.86558E-05	0.013853013
ENSG00000111199.6	TRPV4	2.022580369	1.522405286	0.000343215	0.045176245
ENSG00000163235.11	TGFA	1.995009929	0.894620136	0.000294714	0.040663744
ENSG00000124215.12	CDH26	1.989710791	-1.590508848	0.000381785	0.047752461
ENSG00000091622.11	PITPNM3	1.951802014	-0.538444679	0.000210178	0.03353451
ENSG00000170921.10	TANC2	1.943946558	1.912609494	7.9185E-06	0.004058116
ENSG00000203499.6	FAM83H-AS1	1.927463783	-0.402840899	1.56773E-05	0.005858429
ENSG00000156711.12	MAPK13	1.871209933	2.419680514	0.00021634	0.03354232
ENSG00000163701.14	IL17RE	1.839063602	2.367845065	5.72447E-05	0.015557549
ENSG00000144063.3	MALL	1.828941601	2.466377247	0.000216921	0.03354232

Gene ID	Gene symbol	logFC	AveExpr	P.Value	FDR
ENSG00000125731.8	SH2D3A	1.785113928	0.843924684	7.33064E-05	0.017531965
ENSG00000092295.7	TGM1	1.747368409	-0.406919587	9.45629E-05	0.021202175
ENSG00000171208.5	NETO2	1.726778018	0.964149719	2.49475E-05	0.007990758
ENSG00000063180.4	CA11	1.702846012	0.927555272	5.32228E-05	0.014687028
ENSG00000035115.17	SH3YL1	1.657588388	3.260184342	9.58031E-06	0.00411058
ENSG00000241732.1	RP11-38P22.2	1.616129434	0.339843517	4.8626E-05	0.013853013
ENSG00000103044.6	HAS3	1.592285263	0.698712418	0.000268764	0.038260489
ENSG00000147394.14	ZNF185	1.560404045	1.501000947	9.54207E-06	0.00411058
ENSG00000152454.3	ZNF256	1.556678944	1.076440266	0.000241745	0.035933308
ENSG00000183479.8	TREX2	1.476340338	-1.559628259	0.000305028	0.041765529
ENSG00000135378.3	PRRG4	1.42686905	3.963034543	9.21501E-05	0.02092274
ENSG00000166532.11	RIMKLB	1.379137591	2.715888587	0.000345049	0.045176245
ENSG00000185033.10	SEMA4B	1.304926325	4.770077723	0.00014382	0.026594864
ENSG00000100784.5	RPS6KA5	1.266964189	1.193880795	8.35724E-05	0.019724188
ENSG00000085117.7	CD82	1.255815723	4.207031757	0.00017993	0.029917043
ENSG00000134504.8	KCTD1	1.179070899	1.945419586	0.000115119	0.023184212
ENSG00000178295.10	GEN1	1.140581207	4.258350507	6.12681E-06	0.003270587
ENSG00000138439.10	FAM117B	1.104967344	2.341538734	4.49438E-05	0.013215683
ENSG00000111261.9	MANSC1	1.02450893	3.229147495	5.99806E-05	0.015821634
ENSG00000127337.2	YEATS4	1.000946434	3.214580976	0.000340888	0.045176245
ENSG00000160439.11	RDH13	0.96519217	3.087011461	0.000119746	0.023184212
ENSG00000174106.2	LEMD3	0.835342594	4.127227082	4.83772E-06	0.002799164
ENSG00000139154.10	AEBP2	0.800600401	3.574719878	0.0002424	0.035933308
ENSG00000111596.7	CNOT2	0.70958325	5.616788579	1.90719E-05	0.006417903
ENSG00000158092.2	NCK1	0.691343187	4.471865669	0.000391912	0.048480892
ENSG00000148153.9	INIP	0.647967088	3.838743195	0.000312512	0.042466142
ENSG00000157350.8	ST3GAL2	-0.932809908	3.916522917	0.000221935	0.034024344
ENSG00000116906.7	GNPAT	-0.985813012	6.124558478	0.000116466	0.023184212
ENSG00000168890.9	TMEM150A	-1.13772073	4.783027788	0.000383361	0.047752461
ENSG00000087086.9	FTL	-1.413948289	12.70872926	9.82921E-05	0.021739101
ENSG00000161011.15	SQSTM1	-1.426761585	9.274033463	0.000215223	0.03354232
ENSG00000069869.11	NEDD4	-1.840336847	5.583028022	0.000279181	0.039430496
ENSG00000012504.9	NR1H4	-1.872151225	6.089246877	2.58121E-05	0.008122658
ENSG00000119547.5	ONECUT2	-2.159882242	6.268049627	0.000256755	0.037749331
ENSG00000023839.6	ABCC2	-2.171104017	7.654085697	6.40436E-05	0.016179566
ENSG00000271862.1	RP11-343L5.2	-2.330860996	-0.420943412	0.000161077	0.028105689
ENSG00000174567.7	GOLT1A	-2.364090779	5.646581761	0.000119	0.023184212
ENSG00000132855.4	ANGPTL3	-2.430962222	8.099959614	0.000261294	0.037800136
ENSG00000145217.9	SLC26A1	-2.437547384	3.744907209	0.000265517	0.038100673
ENSG00000135100.13	HNF1A	-2.504602628	5.04023334	1.93213E-05	0.006417903
ENSG00000115363.9	EVA1A	-2.547786615	5.026961409	0.000160277	0.028105689
ENSG00000148935.6	GAS2	-2.568414798	4.094096736	6.13146E-05	0.015939117
ENSG00000084734.4	GCKR	-2.971416007	5.837428684	0.000157976	0.028105689
ENSG00000125798.10	FOXA2	-2.99922049	5.01598045	3.63413E-06	0.002247772
ENSG00000103449.7	SALL1	-3.32420557	5.060734618	1.86127E-05	0.006417903
ENSG00000182902.9	SLC25A18	-3.700948828	5.357351831	0.000348662	0.045318525
ENSG00000111058.3	ACSS3	-4.925855367	5.77690815	2.04556E-06	0.001467645

Table S2. Spearman correlations between the regional mutation frequency of aggregated sample per infection status of PLCs and level of liver H3K4me1 chromatin mark.

Spearman's rank correlation abs (r)	HCC			Mixed			ICC		
	HBV	HCV	NBNC	HBV	HCV	NBNC	HBV	HCV	NBNC
H3K4me1 normal liver	0.869	0.890	0.875	0.628	0.744	0.562	0.361	0.609	0.504
H3K4me1 HBV liver	0.798	0.811	0.800	0.587	0.679	0.514	0.334	0.540	0.452
H3K4me1 HCV liver	0.743	0.752	0.744	0.550	0.635	0.479	0.297	0.493	0.416

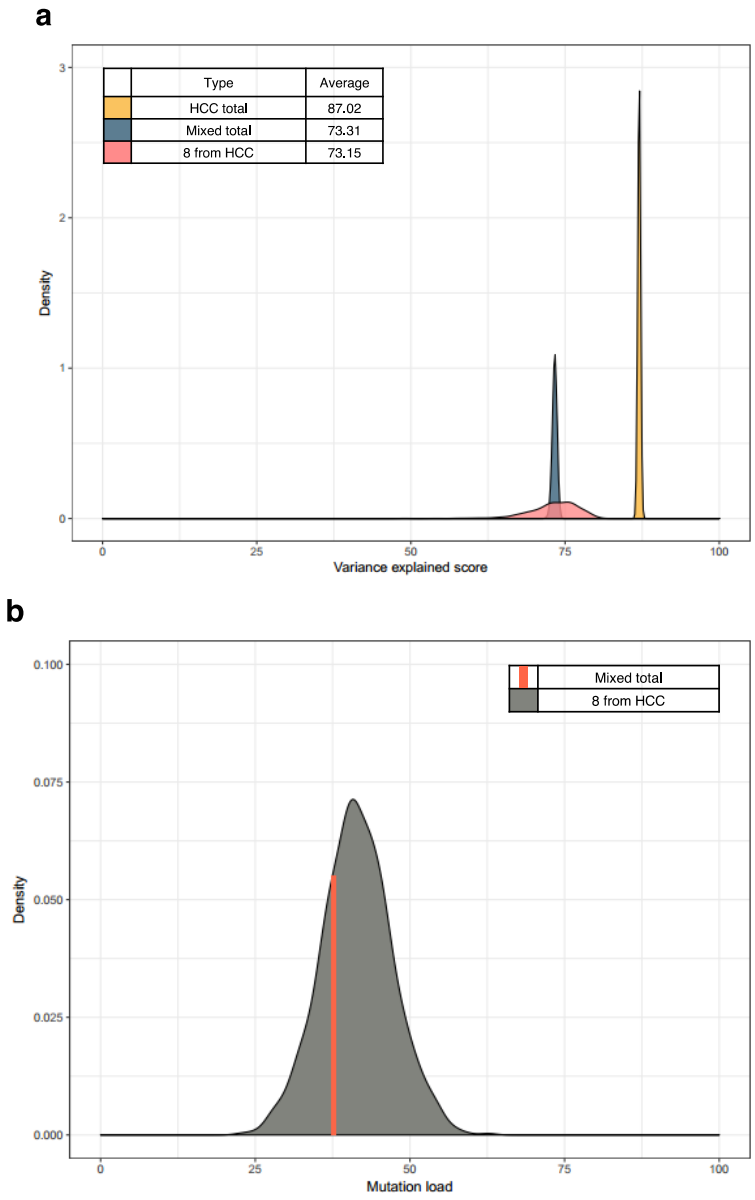

Figure S1. Difference in variance explained scores between the HCC and MIXED type is related to the total number of samples and the aggregated mutation load. (**a**) Distribution of variance explained scores using either all samples or 8 randomly selected samples in 1,000 repeated simulations. Distributions of HCC total (yellow, n = 256) and Mixed total (navy, n = 8) are the result of using all samples for each cancer type. However, pink-colored distribution represents the result of using 8 randomly selected samples in only HCC type. Average variance explained score for each distribution is shown on the top left. (**b**) Distribution of aggregated mutation load at the 1 megabase-level from 8 randomly selected HCC samples in 1,000 repeated simulations. Orange-colored bar represents the aggregated mutation load at the 1 megabase-level from all samples of Mixed type.

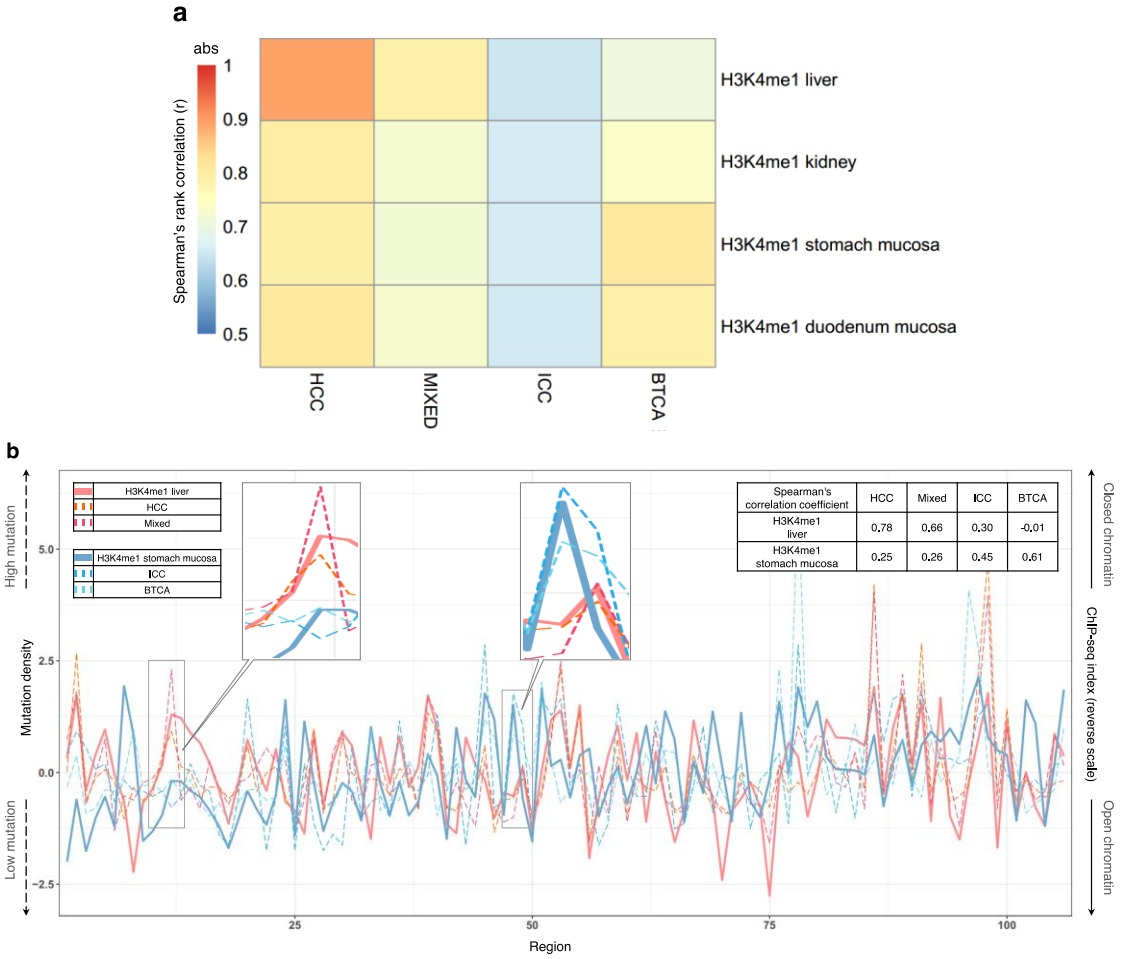

Figure S2. Correlations between cancer genome mutation density and the H3K4me1 chromatin features in different tissue types. (**a**) Heat map with different color depths corresponding to the absolute values of Spearman's ρ statistics. (**b**) Regional mutation density of HCCs, Mixeds, ICCs and BTCAs parallel to the ChIP-seq index (reverse scale) of liver or stomach H3K4me1. Dotted and solid lines represent mutation density and ChIP-seq index, respectively. A total of 106 genomic regions that show top 5% difference from the predicted ChIP-seq count in the regression model between liver and stomach H3K4me1 were selected. Spearman's rank correlations between the mutation density and ChIP-seq index are shown on the top right. Zoomed images are representative regions for cancer type groupings with respect to liver and stomach H3K4me1 level (HCC/Mixed and ICC/BTCA).

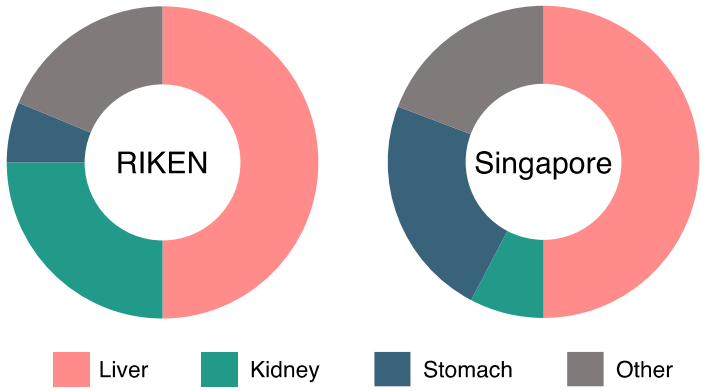

Figure S3. Cell-of-origin prediction distributions for distinct ICC cohorts. Pie graphs represent the percentage of samples getting COO assignments as liver tissue chromatin features (pink), kidney tissue chromatin features (green), stomach tissue chromatin features (navy) or the rest (gray).

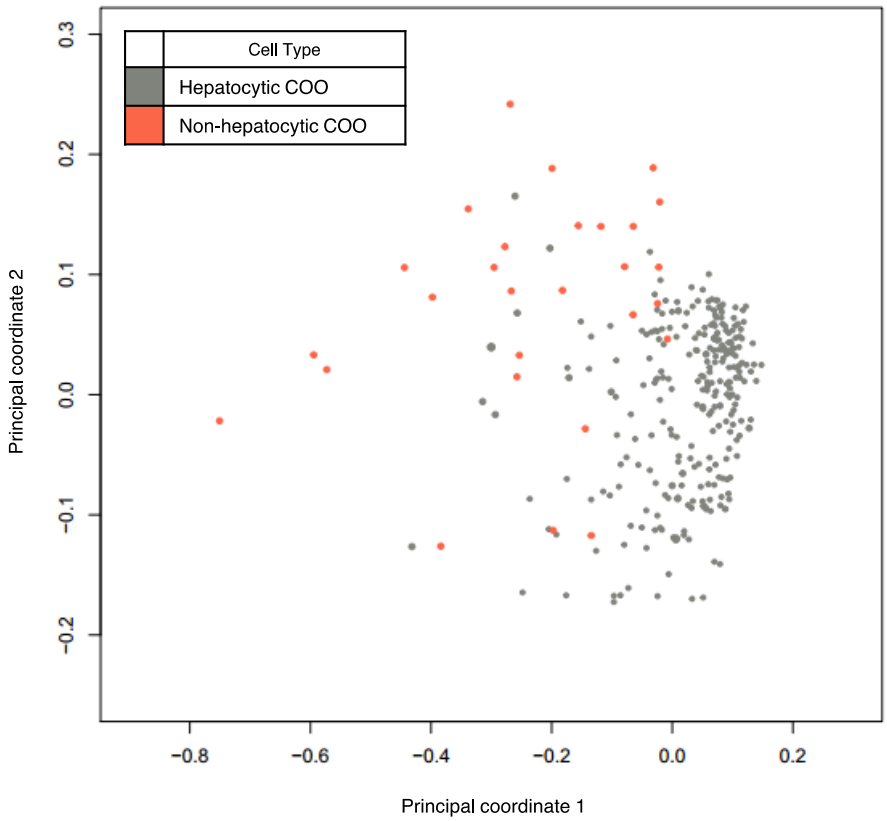

Figure S4. PCOA of individual cancer samples. Hepatocytic COO samples are gray-colored and non-hepatocytic COO samples are orange-colored.

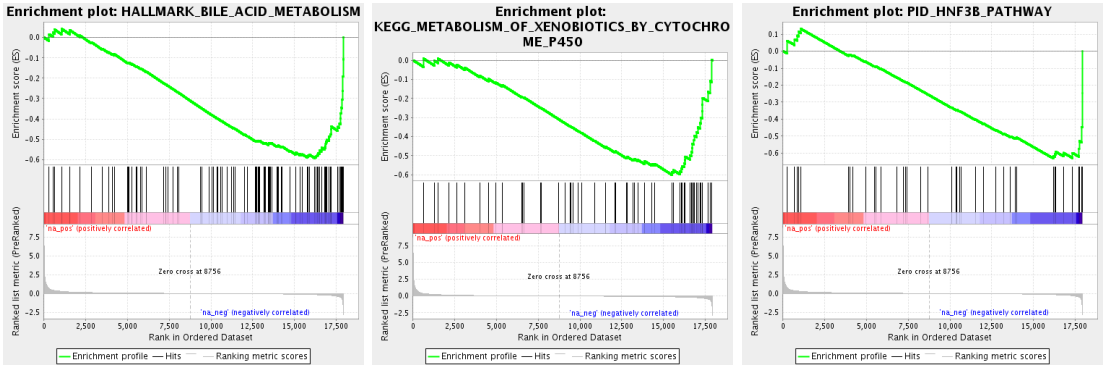

Figure S5. Gene sets that were down-regulated in non-hepatocytic COO HCCs.

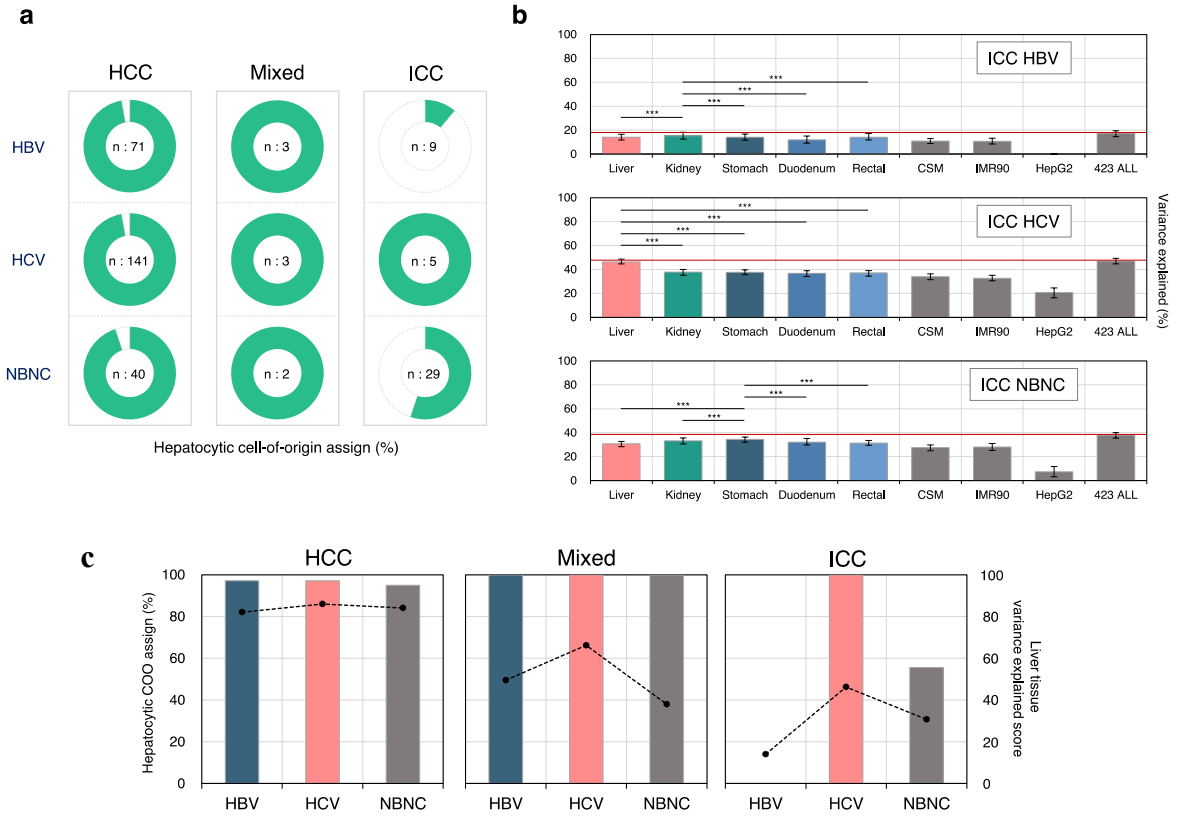

Figure S6. Viral infection status-associated differences in hepatocytic cell-of-origin assignment and variance explained scores. (**a**) Pie graphs represent the percentage of samples assigned as hepatocytic COO (green). The number of samples for analysis is shown at the center of each graph. (**b**) Average variance explained scores for aggregated mutation data depending on the virus type of ICCs are estimated by grouping chromatin features based on each normal cell/tissue type. error bars indicate minimum and maximum scores derived from 1,000 repeated simulations. Red line displays average variance explained score from all 423 epigenomic features. Statistical significance is calculated from tissue with the highest value by using Mann-Whitney U-test (***, $P < 0.001$). (**c**) Bar graphs show the percentage of samples assigned as hepatocytic COO for RIKEN samples. Dots with lines represent average variance explained scores derived by the liver chromatin features.

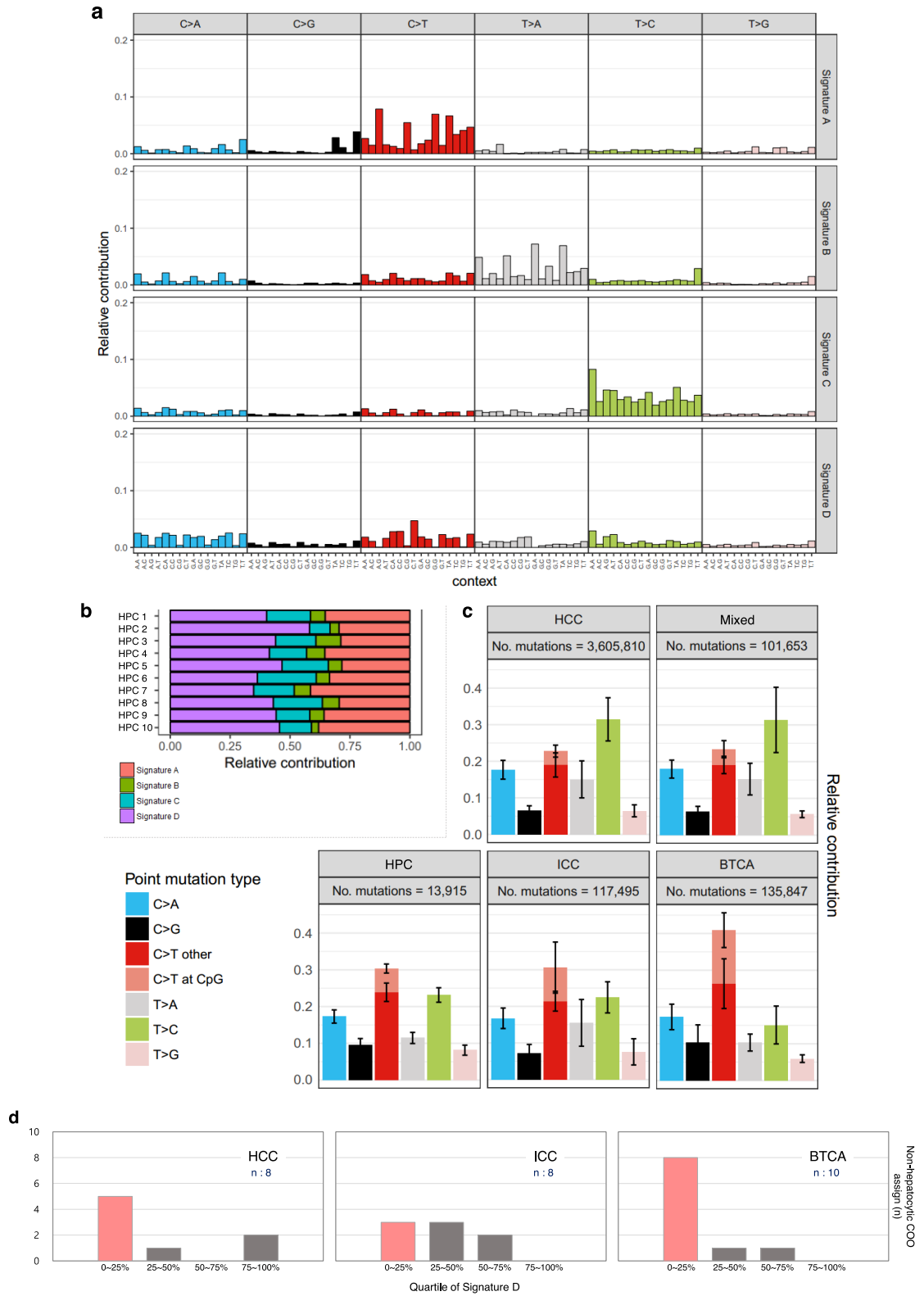

Figure S7. Mutation signature analysis for the genomes of HCC, Mixed, ICC, BTCA-SG and HPC samples. (**a**) Contribution of mutation types to the four mutational signatures derived from the somatic mutations of HCC, Mixed, ICC, BTCA-SG and HPC samples. (**b**) Relative contribution of mutational signatures in each HPC sample. (**c**) Relative contribution of somatic mutation types in each cancer/tissue type. Bar length is calculated as the average relative contribution in each type and error bars show standard deviation. (**d**) Cell-of-origin assignment status based on mutational signatures for HCC, ICC and BTCA. The bar represents the number of non-hepatocytic COO assigned samples with respect to the quartile of signature D contribution. Quartile values are determined by sorting samples of HCCs, Mixed, ICCs, BTCAs and HPCs according to the relative contribution of signature D. The number of samples used in the analysis is shown on each plot.