Heliyon 6 (2020) e03350

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

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

Kyungsik Ha [a], Masashi Fujita [b], Rosa Karlić [c], Sungmin Yang [a], Ruidong Xue [d], Chong Zhang [d], Fan Bai [d], Ning Zhang [d,e], Yujin Hoshida [f], Paz Polak [g], Hidewaki Nakagawa [b], Hong-Gee Kim [a,h,**], Hwajin Lee [a,h,i,*]

[a] Biomedical Knowledge Engineering Laboratory, Seoul National University, Seoul, 08826, South Korea
[b] Laboratory for Cancer Genomics, RIKEN Center for Integrative Medical Sciences, Yokohama, 230-0045, Japan
[c] Bioinformatics Group, Department of Molecular Biology, Division of Biology, Faculty of Science, University of Zagreb, Horvatovac 102a, 10000, Zagreb, Croatia
[d] Biomedical Pioneering Innovation Center (BIOPIC) and Translational Cancer Research Center, School of Life Sciences, First Hospital, Peking University, Beijing, 100871, China
[e] Laboratory of Cancer Cell Biology, Tianjin Medical University Cancer Institute and Hospital, Tianjin, 300060, China
[f] Liver Tumor Translational Research Program, Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center, Dallas, TX, 75390, USA
[g] Department of Oncological Sciences, Icahn School of Medicine at Mount Sinai, 1425 Madison Ave., NY, 10029, USA
[h] Dental Research Institute, Seoul National University, Seoul, 08826, South Korea
[i] Lead contact

ARTICLE INFO

Keywords:
Systems biology
Biocomputational method
Gene mutation
Genomics
Cancer research
Bioinformatics-based prediction of cell-of-origin
Primary liver cancers
Integration of epigenome
Genome and single-cell RNA-Seq data

ABSTRACT

Primary liver tissue cancer types are renowned to display a consistent increase in global disease burden and mortality, thus needing more effective diagnostics and treatments. Yet, integrative research efforts to identify cell-of-origin for these cancers by utilizing human specimen data were poorly established. To this end, we analyzed previously published whole-genome sequencing data for 384 tumor and progenitor tissues along with 423 publicly available normal tissue epigenomic features and single cell RNA-seq data from human livers to assess correlation patterns and extended this information to conduct *in-silico* prediction of the cell-of-origin for primary liver cancer subtypes. Despite mixed histological features, the cell-of-origin for mixed hepatocellular carcinoma/intrahepatic cholangiocarcinoma subtype was predominantly predicted to be hepatocytic origin. Individual sample-level predictions also revealed hepatocytes as one of the major predicted cell-of-origin for intrahepatic cholangiocarcinoma, thus implying trans-differentiation process during cancer progression. Additional analyses on the whole genome sequencing data of hepatic progenitor cells suggest these cells may not be a direct cell-of-origin for liver cancers. These results provide novel insights on the nature and potential contributors of cell-of-origins for primary liver cancers.

1. Introduction

Primary liver cancers (PLCs) are one part of the major cancer types with increasing global disease burden over the years, reaching incidence rates over 900,000 per year (Asrani et al., 2019; GBD Disease and Injury Incidence and Prevalence Collaborators, 2016; GBD Mortality and Causes of Death Collaborators, 2016). The high morbidity and mortality associated with PLCs is due to the complex nature of the disease and the lack of effective diagnostics and treatments besides multi-kinase inhibitors, thus strongly emphasizing the importance of relevant researches on early diagnosis and extensive drug development. In line with this, several research programs endeavored on identifying suitable diagnostic markers and targeted therapy-based treatments for PLCs, including the whole genome and exome-level profiling (Ziogas et al., 2017). Recent comprehensive efforts to investigate the genomics of PLCs have produced novel insights onto 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*). A subset of these mutations are identified as driver mutations and maybe associated with clinical outcomes (Fujimoto et al., 2016; Jusakul et al., 2017). More investigations are underway to fully unveil the mechanisms and processes underlying the progression of PLCs.

* Corresponding author.
** Corresponding author.
E-mail addresses: hgkim@snu.ac.kr (H.-G. Kim), hwajin2k@gmail.com (H. Lee).

https://doi.org/10.1016/j.heliyon.2020.e03350
Received 11 September 2019; Received in revised form 6 January 2020; Accepted 30 January 2020

One of the complex, unanswered questions regarding the progression of PLCs is the nature of cell-of-origins (COOs) corresponding to the various subtypes of PLCs. PLC comprises classical hepatocellular carcinoma (HCC) subtype, which represents ~90% of PLCs, as well as combined hepatocellular and cholangiocarcinoma (cHCC/ICC) and intrahepatic cholangiocarcinoma (ICC), which are the two cancer subtypes displaying biliary phenotype to different extent. The mixed subtype (Mixed), one of the cHCC/ICC subtypes, particularly displays mixed histological features without any clear distinctive boundary between the HCC-like and ICC-like parts, thus posing substantial challenges in inferring the COO for these tumors by either histology or other phenotypic measurements. COOs of PLCs may depend on the location of tumors 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 in animal models proposed possible main 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 (Moeini et al., 2016; Razumilava and Gores, 2014; Sia et al., 2017). None of these have yet been confirmed due to the potential biases accompanied by cell cultures and genetic manipulation-based lineage-tracing animal model systems as well as the lack of human level studies. However, evidences for both differentiated cells and HPCs as the predominant COO for PLCs exist. For example, COOs for HCCs were either reported as solely hepatocytes (Mu et al., 2015) or hepatocytes plus differentiated benign lesions derived from HPCs (Tummala et al., 2017). As the COOs for ICCs, either hepatocytes which underwent conversion into cholangiocytes (Sekiya and Suzuki, 2012) or the biliary epithelial cells themselves (Guest et al., 2014) 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 or trans-differentiation of hepatocytes (Mu et al., 2015) and cholangiocytes (Raven et al., 2017; Russell et al., 2019) after the liver injury as potential sources of progenitor cells and PLCs, which further enhances the complexity for the identification of liver cancer COOs. Efforts to extrapolate these COO-related complexities by utilizing actual human cancer tissue data itself are scarce with one article partly visiting this issue at a limited sample-level (Wardell et al., 2018), and no studies were yet performed in a comprehensive, inter-cohort manner.

Here, we performed a computational approach to dissect out the putative COOs on each cancer subtype of PLCs and interrogated possible individual tumor-level heterogeneity in COOs. For this, we analyzed the whole genome sequencing data from 341 of PLCs (256 HCCs, 29 Mixed, and 56 ICCs) and 12 extrahepatic biliary tract cholangiocarcinoma samples (BTCAs) based on the assumption that this cancer type would predominantly display cholangiocytic COO, and 10 of HPCs to assess the possibility of being a common COO for PLCs, along with 423 chromatin features at the epigenome-level (methods). Since chromatin marks were generated from tissue-level samples, we attempted to complement our findings on the correlations between somatic mutation landscape and chromatin features by utilizing single cell RNA-seq (scRNA-seq) data derived from human liver tissue (MacParland et al., 2018) to dissect out the relationships between the gene expression features from normal liver cell types and somatic mutation landscape of PLCs. 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.

2. Results

2.1. Aggregate sample-level correlations between chromatin marks and somatic mutations of PLCs

Based on the previous findings about the associations between the chromatin feature levels and regional variations in somatic mutation frequencies of tumors (Polak et al., 2014, 2015) and applying this knowledge onto machine-learning based COO predictions on several cancer types (Kübler et al., 2019), we first hypothesized that the whole-genome mutation landscape of hepatocytic PLC subtype (HCCs) would exhibit a closer relationship with liver tissue (surrogate tissue for hepatocytes) chromatin marks, whereas the mutation landscape of partial or fully biliary PLC subtypes (Mixed and ICCs) and the BTCAs would likely to display stronger 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 phenotypes and anatomical locations. 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 that explained the possible variances in regional somatic mutation frequencies. To conduct the analysis, we utilized somatic mutation frequency data at a 1-megabase window 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 displaying significance for HCCs, and a stomach tissue chromatin mark served as the first-rank feature for ICCs and BTCAs ($P < 2.2e\text{-}16$, Mann-Whitney U-test between the first and second rank features of each PLC subtype; Figure 1a). Surprisingly, liver tissue chromatin marks were major features explaining the regional mutation variation of Mixed subtype. This result indicates a possible tendency of putative COO towards to the hepatocytes for the Mixed subtype, albeit known molecular heterogeneity among individual tumors (Moeini et al., 2017) and the partial biliary phenotypes in histology. The overall lower variance explained scores for Mixed and ICCs compared to the HCCs were at least in part likely due to the lower number of the samples and the total mutation load (Figure S1a, b), indicating that the actual correlation between the liver tissue chromatin features and the somatic mutation landscape of Mixed may be similar to that of HCCs. In line with these results, spearman correlations between the regional mutation frequency of HCCs or Mixed and liver H3K4me1 chromatin mark level was the largest when comparing to different chromatin marks from a possible pool of surrogate tissues, whereas stomach H3K4me1 chromatin mark level showed the highest correlation with the regional mutation frequency of BTCAs (Figure S2a). Spearman correlation values among the regional mutation frequency of ICCs and H3K4me1 of different tissues were overall low without displaying any tissue type dependent differences, which can be due to both the lower mutation load of ICCs and the possible intrinsic COO heterogeneity. These correlation patterns were more exemplified when sub-setting the genomic regions according to the top 5% difference in ChIP-seq counts between liver and stomach H3K4me1 marks (Figure S2b). Similar to the spearman correlation results, the regional quintile-based mean mutation density data of HCCs and Mixed showed relatively higher association with the liver tissue H3K4me1 level comparing to the stomach tissue H3K4me1 level, while the mean mutation data for ICCs and BTCAs displayed higher association towards the stomach tissue H3K4me1, with ICCs as a lesser extent (Figure 1b). Collectively, these results demonstrate that COO-associated chromatin features can delineate the relationships with the mutation landscape of PLCs and BTCAs.

2.2. Aggregate sample-level correlations between single cell RNA-seq data and somatic mutations in PLCs

Previous publication showed that gene expression data can explain regional somatic mutation variance, albeit at a lower level compared with the chromatin features (Polak et al., 2015). As with any major tissue types, liver tissue contains multiple cell subpopulations including hepatocytes, cholangiocytes, stellate cells and other rare cell types, which suggests a potential limitation of mixed cell subpopulations when using traditional bulk tissue-level RNA-seq data in such analysis. In our study, we revisited correlation levels between gene expression and the

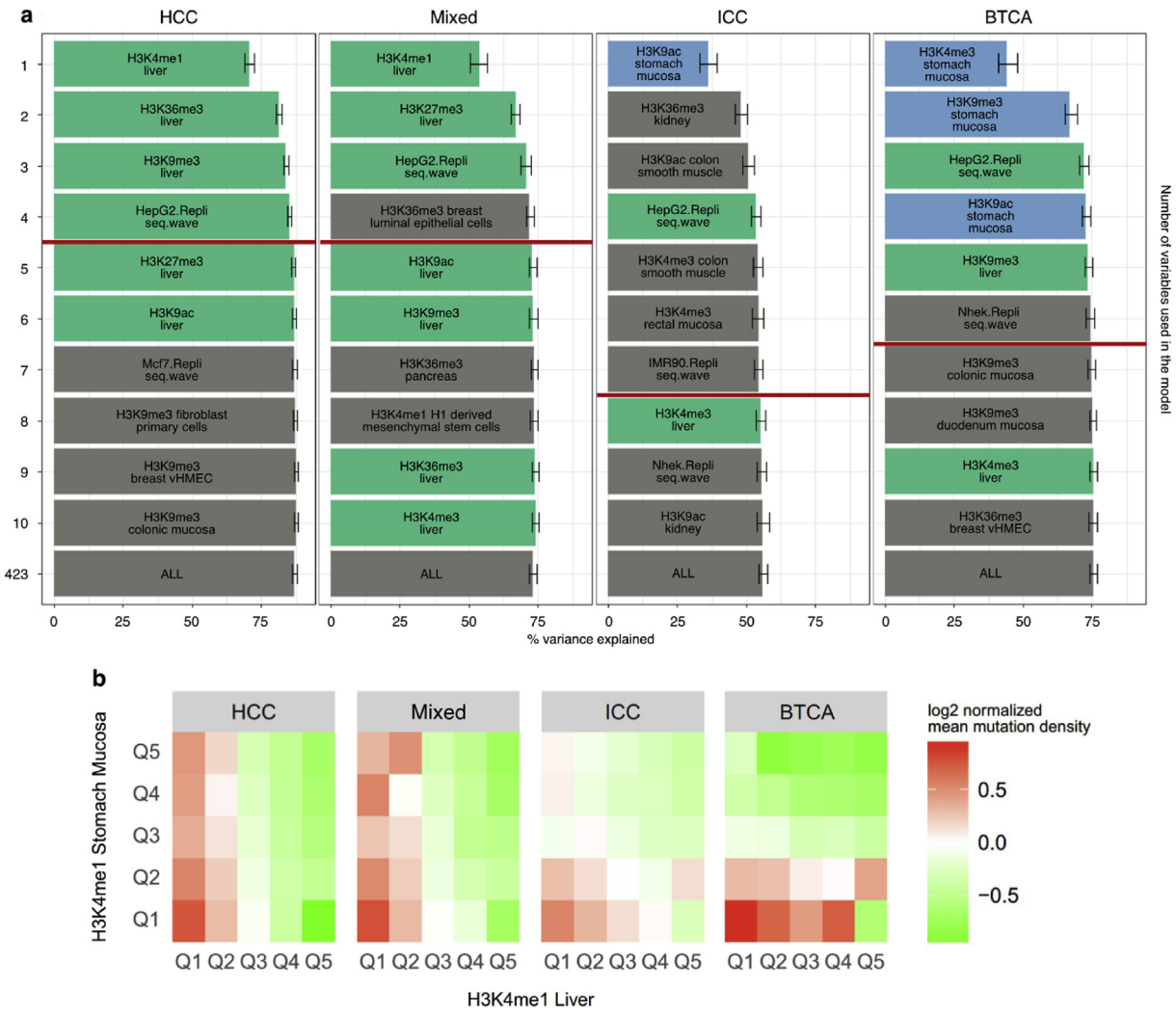

Figure 1. Cell-of-origin chromatin features delineating relations with the regional mutation frequency of HCCs, Mixed, ICCs and BTCAs. (a) Random forest regression-based chromatin feature selection using aggregated somatic mutation frequency data from HCC, Mixed, ICC and BTCA-SG samples. The rank of each chromatin feature was determined by importance values. Bar length represents the variance explained scores, and the 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 standard error of the mean. 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.

somatic mutation landscape for PLCs by utilizing recently published human liver scRNA-seq data (MacParland et al., 2018), thus taking into account the heterogenous cell types within a liver tissue. After sub-selecting four cell clusters representing hepatocytes and one cluster corresponding to cholangiocytes (methods), we first assessed the relationship between gene expression features and somatic mutation landscape of PLCs for all of the 1-megabase genomic regions after employing a single-cell-level RNA transcript detection rate (DR) threshold on gene expression data (methods). Spearman correlation values between either DR or mean detected transcript count level (MDTC) and somatic mutation frequencies for PLC subtypes showed significant but generally lower correlation values than when using H3K4me1 chromatin features (spearman coefficient (absolute value) $<$ 0.52 for HCC, $<$ 0.45 for Mixed, $<$ 0.32 for ICC and $<$0.45 for BTCA). We next used the top 5% difference in H3K4me1 ChIP-seq counts between liver and stomach tissues, which are the most representative regions used in the previous analysis showing differences in correlations between regional somatic mutation frequencies for PLCs and chromatin features. Results assessing the correlation between the H3K4me1 chromatin features and DR or MDTC for these sub-selected regions revealed that the DR values were more representative of demonstrating expected correlations with chromatin features for both tissue types (Figure S3a). A subsequent analysis was conducted to assess the correlations between DR values from either hepatocyte or cholangiocyte clusters and regional somatic mutation variations of PLCs in the subset regions. Results showed that although the correlation coefficients derived from DR values were less robust than the chromatin features, (consistent with the previous report (Polak et al., 2015)), the observed correlation tendencies were similar, especially for the somatic mutation landscapes for ICCs and BTCAs. (Figure S3b).

Based on the results above, we next examined the possibility of using DR value features from individual liver cell types by conducting random-forest feature selection method (methods). Although showing lower variance explained scores, our results displayed consistencies with the chromatin-based feature selection results (Figure 1a) by showing hepatocyte DR feature as the first rank for HCCs and Mixed, and cholangiocyte DR feature as the first rank for ICCs and BTCAs (Figure S3c).

Collectively, our results using DR gene expression feature complemented the chromatin feature-based aggregate-level analyses and further confirmed the relationship between the molecular features derived from the putative COO and regional somatic mutation frequencies of PLCs.

2.3. Individual sample-level cell-of-origin predictions

To further assess the differential mutation landscapes and possible COOs for PLCs and BTCAs at the individual sample level, we conducted a random forest algorithm-based COO analysis for each sample (methods). This individual sample-based COO analysis demonstrated the dominance of a hepatocytic predicted COO for HCCs, in contrast to the predictions for BTCAs, which showed stomach tissues (a proxy tissue for extrahepatic cholangiocytes) as a major putative COO (Figure 2a). For the mixed subtype, hepatocytic COO was solely predicted for the 8 samples that were used for the aggregate sample-level random forest analysis. This result was replicated for an additional 20 Mixed subtype samples from another cohort (Xue et al., 2019) (Figure S4a), which is yet again in line with the aggregate-level correlation results and the recent publication on the monoclonal origin of mixed subtypes enriched with HCC-like gene expression-level features (Xue et al., 2019). For ICCs, however, both hepatocytes and proxy tissues for cholangiocytes (kidney and stomach) were predicted to be possible major COOs. This COO prediction pattern was consistent between different ICC cohorts (Figure S4b), thus emphasizing the consistent heterogeneity of COOs and inferring that the somatic mutation landscape can harbor the signature of cell type trans-differentiations and plasticity involved in liver injury (Monga, 2019), which is most likely to occur prior to the development of ICCs. Our results not only replicated earlier findings on the COOs of HCCs, ICCs and extrahepatic distal cholangiocarcinoma (DCCs) (Wardell et al., 2018), but also adding a couple of novel aspects including 1) the complete predominance of hepatocytic predicted COO for Mixed tumors (28/28) and 2) 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, six HCC samples showed non-hepatocytic predicted COO, thus implying a possibly distinct COO for a subset of HCCs that may be linked to differential tumor pathology. Overall, our results suggest that the predominant COO for the HCCs and Mixed would most likely to be hepatocytes. Also, our evidences point to the cholangiocytes as the likely predominant COO for BTCAs, whereas the COOs of ICCs tend to vary by individual samples. These results confirm the importance of anatomical locations on the COOs of PLCs and BTCAs.

Next, we utilized DR gene expression features derived from human liver tissue as an alternative to chromatin features from liver, kidney and stomach tissues. Application of DR features from a total of 20 scRNA-seq clusters for random forest-based COO prediction (methods) to 20 Mixed subtype samples with positive variance explained scores cross-confirmed the chromatin feature-based COO prediction results (18 out of 20 showing hepatocytic COO; Figure S5). For ICCs, only 5 out of 56 samples displayed positive variance explained scores, further implicating

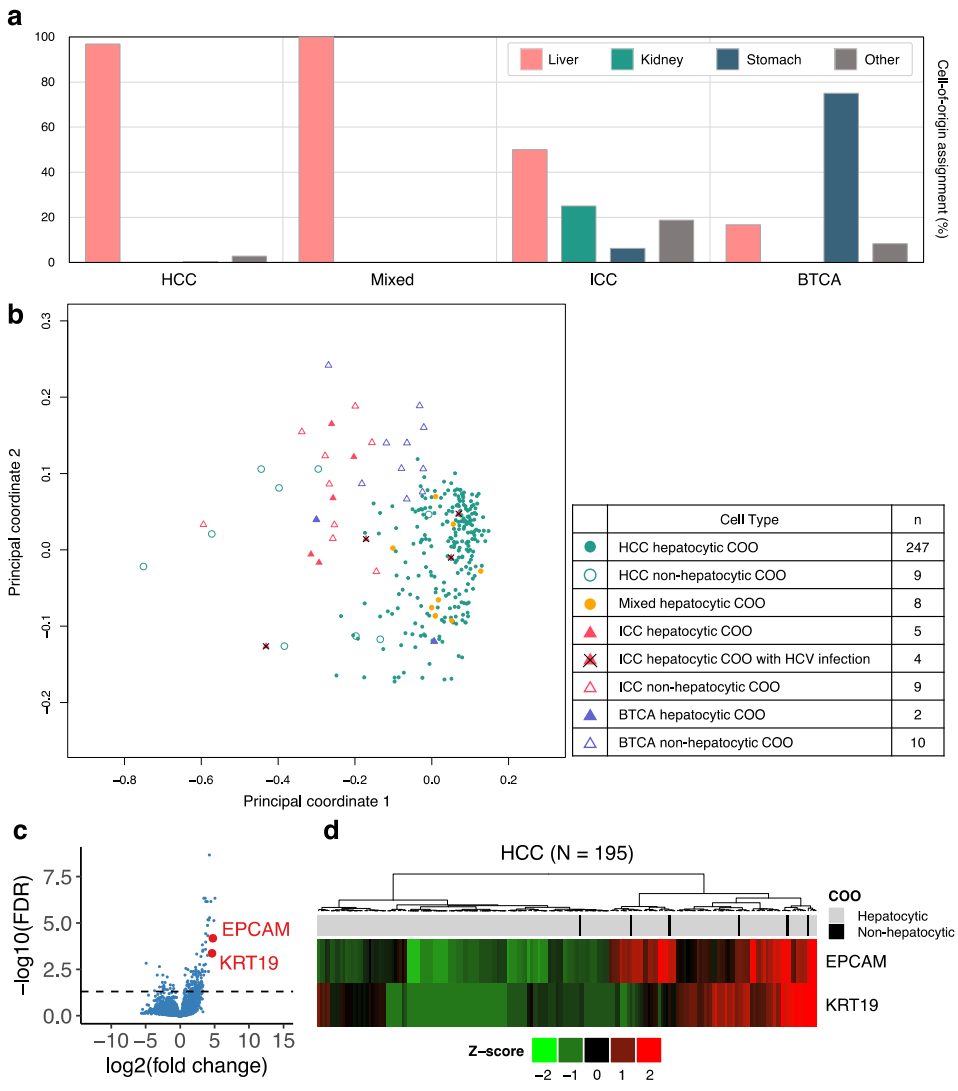

Figure 2. Analysis of COOs for individual cancer samples. (a) Prediction of COO via grouping of chromatin features for each normal tissue type. The bar graph depicts the percentage of samples with respect to the assigned COO by liver tissue chromatin features (pink), kidney tissue chromatin features (green), stomach tissue chromatin features (navy) or the rest (gray). (b) Principal coordinate analysis 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.

chromatin features as better predictors of regional somatic mutation frequencies compared with the scRNA-seq based gene expression features. This result is also in line with the aggregate-sample level correlation results discussed earlier.

Along with these results, principle coordinate analysis (PCOA) result revealed that the PLC samples with hepatocytic predicted COO tend to aggregate as a cluster, displaying principle coordinate 1 value over 0 (Figure S6). In terms of 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 (Figure 2b), reflecting the distinct mutation landscape patterns.

To demonstrate whether HCCs with non-hepatocytic predicted COO have a unique gene expression patterns compared with the hepatocytic predicted HCCs, we analyzed the genome-wide gene expression profiles. Among the non-hepatocytic- and hepatocytic predicted HCC samples, tumor RNA-seq data were available for 6 and 189 samples, respectively (Fujimoto et al., 2016). A comparison of gene expression levels between them showed that 124 genes were up-regulated and 21 were down-regulated in non-liver-origin HCCs (FDR <0.05, absolute logFC >0.647; Table S1). Interestingly, the upregulated genes included an epithelial cell marker *EPCAM* and a cholangiocyte-specific marker *KRT19* (Figure 2c). Clustering analysis confirmed that HCCs with non-hepatocytic predicted COO were enriched in a cluster that expressed more *EPCAM* and *KRT19* (Figure 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 (Figure S7). This result indicates that the functional similarity to hepatocytes is being less observed in HCCs with non-hepatocytic predicted COO. Collectively, the mRNA expression in non-hepatocytic predicted HCCs partly resembled that of biliary epithelial cells, which follows the preceding publication about EPCAM-positive ductal cells as a possible COO for HCCs at an inflamed condition (Matsumoto et al., 2017). We also compared hepatocytic- and non-hepatocytic predicted HCCs in terms of clinical features (including tumor stage and survival), but we found no statistically significant difference in these features, which suggest that the COO assignments for HCCs may be independent of the clinical prognosis.

Hepatitis virus infections in ICCs have been previously reported to display distinct clinical features and prognoses depending on which virus type is infected (Wang et al., 2016). Since a prior publication described the association between hepatitis virus infection status and liver COO assignments without any subgrouping of virus types (Wardell et al., 2018), we tested whether there are any hepatitis virus-type specific tendencies in COO predictions and the variance explained scores for the somatic mutation landscape of PLCs. Upon grouping PLCs with hepatitis B virus (HBV) and hepatitis C virus (HCV) infection status, our analysis revealed that HCCs and Mixed samples were assigned primarily to hepatocytic COO regardless of the hepatitis virus infection status. In contrast, ICC samples displayed differential COO predictions based on the viral infection status of the patients. In the case of HCV-infected ICCs, all of the samples showed hepatocytic predicted COO (binomial probability of 0.08, two tailed), whereas HBV infected ICCs were predominantly predicted as non-hepatocytic COO (8 out of 9, binomial probability of 0.04, two tailed) (Figure S8a, c). These results might reflect differential effects of viral infections onto different cell types within the liver tissue and their progression into ICCs, which would depend on the hepatitis virus types. This implication is consistent with the previous reports on differential ability of HBVs (positive infectivity) and HCVs (negative infectivity) to infect cholangiocytes (Blum et al., 1983; Fletcher et al., 2015). Furthermore, spearman correlation values between the regional mutation frequency of aggregated samples grouped by HBV or HCV infection status and normal liver tissue H3K4me1 chromatin mark level was higher for HCV-infected ICCs compared with any other ICCs with different virus infection status, and this result was fully replicated when using H3K4me1 chromatin marks derived from HBV or HCV-infected liver tissues, thus providing additional evidences (Table S2). In line with these results, and using variance explained scores for the ICCs calculated by using a total of 9 cell or tissue types, we discovered that chromatin features with the highest level of variance explained scores were derived from different tissues depending on the hepatitis infection status of ICCs (HBV = kidney tissue, HCV = liver tissue, NBNC = stomach tissue) (Figure S8b). Although a limited number of virus-infected ICC samples, our results indicate a potential skewness of COO of ICCs depending on the virus infection status, and a separate cohort level study with larger number of samples is strongly warranted.

2.4. Hepatic progenitor cells as a possible cell-of-origin for PLCs

EPCAM-positive HPCs, so called as oval cells, are a progenitor cell type located inside the Canal of Hering. HPCs harbor differentiation capacity into both hepatocytes and cholangiocytes, and also have been suspected to be a possible COO for PLCs. To examine the possibility of HPCs as a possible COO for different subtypes of PLCs, we performed random forest feature selection analysis using somatic mutation frequency data for HPCs (Blokzijl et al., 2016) at an aggregate sample level. Results from this analysis demonstrated that the mutation landscape of HPCs cannot be explained adequately by the normal tissue chromatin landscape, with negative-value variance explained score for the top 1st rank chromatin feature and 25% for the total 423 chromatin features (Figure 3a). To check whether the results from HPCs were due to the lower mutation load or possible differences in mutation accumulation patterns intrinsic to the adult stem cells, we utilized the mutation landscape data of colon stem cells (Blokzijl et al., 2016). Aggregate sample level random forest feature selection analysis of colon stem cells displayed variance explained score greater than 40% for the H3K9me3 rectal mucosa chromatin mark and above 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 ranked feature, with greater than 28% variance explained score, implying that either the lower mutation load or the stem cell specific mutation accumulation patterns might not be a contributing factor for the feature selection analysis results from two different adult stem cells. These results also infer distinct mutation landscape between the HPCs and PLCs through differential variance explained score patterns, thus suggesting that HPCs might not be a direct COO of PLCs.

2.5. Relationship between mutation signatures and COO predictions

Previous evaluation on the mutation signature of HPCs identified a specific age-associated mutation signature displaying a correlation with replication timing and average chromatin levels of cell lines registered in the ENCODE project (Blokzijl et al., 2016). Based on these findings, we conducted mutation signature analysis on the HPCs along with the PLCs and BTCAs to discover any relationship between the mutation signature proportions and COO assignments. 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 (Figure S9a-c). Next, we assessed whether the proportion of signature D correlates with COO assignment for PLCs. As demonstrated in Figure 3b, the relative contribution of signature D was significantly lower for non-hepatocytic predicted HCCs and ICCs comparing to the hepatocytic-predicted HCCs/ICCs and all of the HPCs. Moreover, several evidences point out that the correlation between the relative proportion of the mutation signature and the COO assignment was specific and consistent for signature D. One is that the proportion of the other three signatures (A, B and C) was not significantly associated with the COO assignments for ICCs ($P > 0.57$), and two signatures (A, B) showed no significant associations with the COO assignments for HCCs ($P > 0.24$). Also, the mutation type patterns of HPCs were more comparable to those of ICCs and BTCAs rather than the HCCs and Mixed, in contrast to the

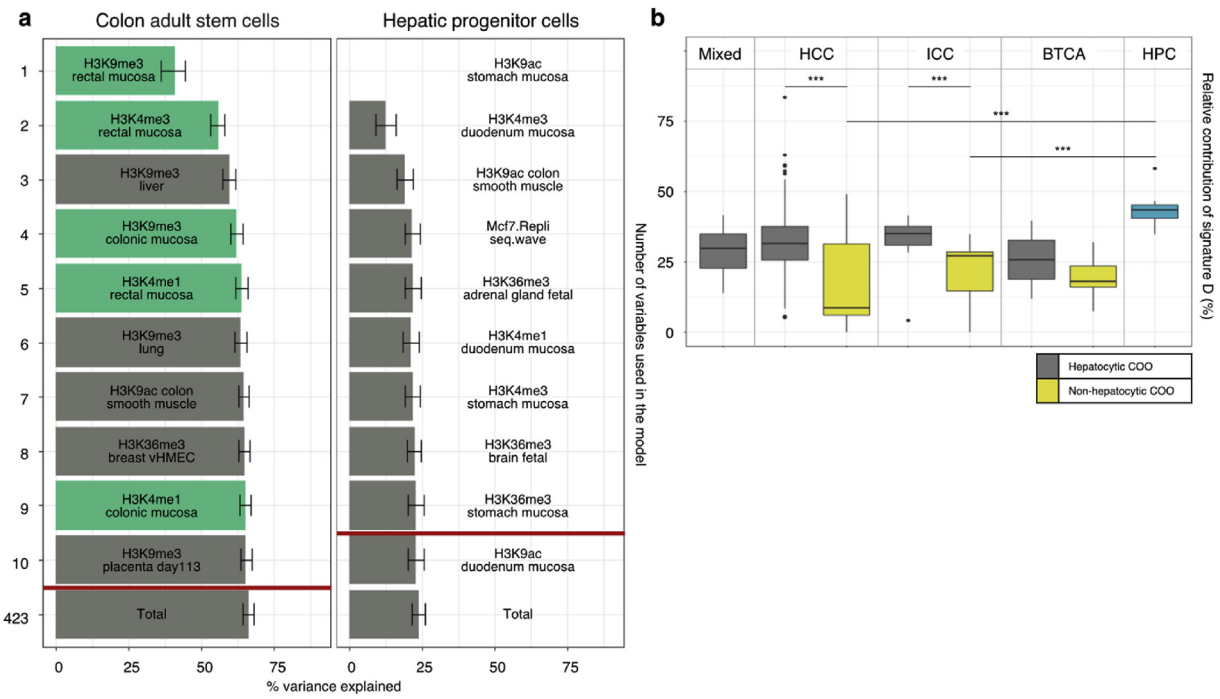

Figure 3. Hepatic progenitor cells display distinct mutation landscape and mutational signature processes compared to the genomes of PLCs. (a) Chromatin feature selection in relation to the regional mutation frequency of colon adult stem cells and hepatic progenitor cells. The chromatin features related to each tissue type are green-colored. (b) The box plot shows 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 was calculated by using a Mann-Whitney U-test (***, $P < 0.05$). BTCAs were excluded from the statistical analysis because only two samples were predicted as hepatocytic COO.

findings on the skewness of COO assignment depending on the signature D status. Furthermore, major proportion of the non-hepatocytic predicted COO samples were located in the lower quartile for the signature D proportions (Figure S9d). Collectively, these results provide a novel perspective with respect to the importance of age-associated mutation signature levels on COO assignment, and thus reflect the distinct mutation landscapes between hepatocytic and non-hepatocytic predicted COO samples.

3. Discussion

In this paper, we applied random-forest machine learning algorithm and other computational analyses to whole genome sequencing data of PLCs and epigenomics data/scRNA-seq data derived from normal tissues to elucidate unique association patterns between the two features and identify possible COO distribution for PLCs at the subtype and individual tumor tissue level. Results from these analyses would help to understand the complex and heterogeneous nature of liver cancer COOs and the contribution of chromatin marks on differential regional somatic mutation landscapes during the progression of 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 (Ha et al., 2017; Polak et al., 2014, 2015). Consistent with this idea, our results show that 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 liver tissue and the top correlative chromatin marks for 12 of BTCAs were from the stomach tissue, which are also concordant to the previous studies on HCCs and DCCs (Wardell et al., 2018). Also, analysis of the scRNA-seq data from human liver tissue complemented the chromatin feature-based data by using DR value feature data from the actual cell types inside the liver tissue. To note, a lower level of variance explained scores were observed for ICCs comparing to any other PLC subtypes, using either chromatin features or the DR value features. We speculate that the potential contributor to these differences in variance explained scores might be either 1) lower mutation load or 2) the higher level of heterogeneity in COOs.

Genetically-engineered mouse model (GEMM) lineage tracing studies reported COO-dependent discrepancies with respect to the oncogenic alterations at the molecular level (Vicent et al., 2019). In the case of ICCs, mouse models either utilizing thioacetamide administration or Trp53 genetic loss can direct different cell types (hepatocytes vs cholangiocytes) into ICCs with concomitant Notch signaling activation (Guest et al., 2014; Sekiya and Suzuki, 2012). For HCCs, most of the mouse models revealed that this cancer subtype mainly originates from hepatocytes, but the emergence of HPC-derived benign lesions could be identified in conjunction with galectin-3 and α-ketoglutarate paracrine signals (Tummala et al., 2017). Our COO prediction results not only do conform with these reports but also stress out the importance of further large cohort-level investigation on the major COOs of each subtype of PLCs and the potential COO variability, especially in the context of distinct or co-existing molecular alterations. Altogether, these researches would remain highly necessary for a better understanding of the cancer progression for PLCs along with the early-stage diagnosis and the treatment selection.

Several publications provided pieces of evidence on the injury-mediated plasticity of hepatocytes by demonstrating the ability to transdifferentiate into cholangiocytes (Michalopoulos et al., 2005; Sekiya and Suzuki, 2014; Yanger et al., 2013) at in vitro and/or in vivo. Moreover, several lines of lineage-tracing based evidence show that the transdifferentiated hepatocytes can arise ICCs indifferent mouse models (Fan et al., 2012; Sekiya and Suzuki, 2012; Wang et al., 2018). These transdifferentiation processes are governed mainly by the activation of Notch1/2 and Akt signaling, which is renowned to be crucial for the formation of ICCs at least in part by direct transcription and overexpression of cyclin E gene (Zender et al., 2013). Consistent with these observations, our random forest-based COO predictions also point out the

possibility that the hepatocytes are indeed one of the major COOs of ICCs, alongside with the cholangiocytes. These results implicate that the somatic mutation landscape of tumors can harbor the information about the history of cancer initiation and progression, which may enable to detect the potential cellular transdifferentiation during the course of cancer development and accompanied somatic mutation accumulations.

The COOs for PLCs were a subject of debate for a number of years, not only due to the discovery of several types of HPCs (Cardinale et al., 2011; Wang et al., 2015), but also to the facultative regeneration of hepatocytes and cholangiocytes displaying trans-differentiation, which mainly occurs during the inflammation or liver injury (Mu et al., 2015; Raven et al., 2017). Our prediction results, at least, favor differentiated cells rather than progenitor or stem cells as origins for PLCs. This conclusion is based on the findings that 1) normal liver (representing hepatocytes), kidney, and stomach (surrogate for the cholangiocytes) tissues can mostly explain the COO of PLCs, and 2) the somatic mutation profile of HPCs is not adequately explained (variance explained score <24.04) by the normal tissue chromatin marks. 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 imply a distinct 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 happen in 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 could serve as 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.

In summary, our results on the COO of PLCs discovered several novel aspects of COO distribution in different PLC subtypes. We believe that these results not only validate the *in vitro* and *in vivo* data from previous publications on COOs of PLCs through human data but also address some new aspects of individual-level differences in tumor biology and clinical pathology of PLCs, and provide a robust and relevant way of studying cancer COOs in a human system. Ultimately, our results support arguments for the necessity of personalized medicine for cancer treatments, combined with genomics and other molecular signatures.

3.1. Limitations of the study

In this study, our current standpoints for the limitations are as follows: 1) Some of our results were derived by using surrogate tissue chromatin marks rather than the true intrahepatic cholangiocyte chromatin marks. Due to the technical limitations in collecting sufficient amounts of pure intrahepatic cholangiocytes for ChIP-seq assay, we focused our efforts to complement this issue by employing scRNA-seq data. 2) Utilizing chromatin mark data from HPCs would have been ideal, but technical limitations were present on purifying these cell populations from human liver tissues with sufficient amounts for ChIP-seq profiling. In our study, we instead analyzed the somatic mutation landscape of HPCs to assess the somatic mutation accumulation patterns and mutation signatures. 3) The number of samples for the HCV-infected ICCs was too low to derive a concrete conclusion, and this was due to the scarcity of ICC populations with HCV infection. We believe that our study supplies scientific rationale for a further study on differential properties of the somatic mutation landscape in hepatitis virus-infected ICCs.

4. Methods

4.1. Preprint publication

The article was previously published as a preprint in bioRxiv (Ha et al., 2019).

4.2. 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 from 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 consisting 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) (Cibulskis et al., 2013). In addition to the data sets listed above, WGS-derived somatic mutation profile from additional 31 stem/progenitor samples (10 HPCs and 21 colon adult stem cells) and 38 ICCs from previous studies (Blokzijl et al., 2016; Jusakul et al., 2017) were utilized for the analysis related to hepatic progenitor cells (Figure 3 and Figure S9) or as an independent cohort for predicting the COO of ICCs (Figure S4b) and assessing viral-infection associated COO predictions for ICCs (Figure S8a). Furthermore, additional WGS data from 21 Mixed subtype samples from a recent study (Xue et al., 2019) were also used for the COO prediction as another independent cohort. 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 (Consortium, 2012) and the NIH Roadmap Epigenomics Mapping Consortium (Roadmap Epigenomics et al., 2015). NIH Roadmap epigenomics data can be accessed through the NCBI GSE18927 in the Gene Expression Omnibus site (https://www.ncbi.nlm.nih.gov/geo/). Chromatin data for liver tissues derived from hepatitis virus infected patients (donor HPC8 and HPC17) were obtained from the IHEC (https://epigenomesportal.ca/ihec/download.html). All chromatin data applied to this study derived from a post normalized bed file. The epigenomics data were used in previous studies for assessing the relationship between chromatin marks and somatic mutation landscape of tumor (Polak et al., 2015) and pre-cancerous lesion-tumor pairs (Ha et al., 2017). 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. Subsequently, 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 were also used to calculate DNase I hypersensitivity and Repli-seq profiles in each 1-megabase region. All these calculations were performed using BEDOPS (Neph et al., 2012).

4.3. Principal coordinate analysis

PCOA was employed to represent the 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.

4.4. Feature selection based on random forest algorithm

Our feature selection analysis applied a modified version used in the previous study (Polak et al., 2015). 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.

4.5. 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 (Figure 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 importance ranking among variables from 423 chromatin features due to the fact that liver chromatin features were the only major type in the aggregated feature selection results for HCCs. For our purpose, we considered the samples with positive variance explained score as relevant samples for the COO assignments.

4.6. Signature analysis of mutational processes

A nonnegative matrix factorization (NMF) algorithm was employed to investigate mutation signatures as described in the previous study (Blokzijl et al., 2018). This methodology was utilized by factoring out the frequency matrix of 96-trinucleotide mutation contexts from HCC, Mixed, ICC, BTCA-SG and HPC samples.

4.7. Gene expression analysis

RNA-Seq experiments of HCC samples were performed previously (Fujimoto et al., 2016), and the data has been deposited in the European Genome-phenome Archive. The reads were aligned onto the reference human genome GRCh37 using TopHat v2.1.1. Raw read counts per gene were computed using HTSeq with the GENCODE v19 annotation. Differential gene expression between hepatocytic- and non-hepatocytic-origin HCCs was analyzed using limma-voom v3.26.9 (Ritchie et al., 2015). Gene set enrichment analysis (GSEA) was performed using the GSEAPreranked v5 module on the GenePattern server (https://genepattern.broadinstitute.org).

Assessment of relationship between aggregate sample-level somatic mutation landscape and Single-cell RNA-sequencing (scRNA-seq) data.

Data acquirement from single cell clusters was performed by running scClustViz algorithm (Innes and Bader, 2018) on previously generated human liver scRNA-seq data (MacParland et al., 2018). Two central venous hepatocyte clusters (Cluster 1 and 3), two periportal-like hepatocyte clusters (Cluster 5 and 14) and one cholangiocyte cluster (Cluster 17) was selected as representative cell clusters for this analysis. Spearman correlation level association was assessed between either of the two gene expression factors (within-cluster level cellular transcript detection rate, DR; mean detected transcript count for the cells harboring detectable transcript level, MDTC) (Innes and Bader, 2018) derived from representative clusters and chromatin features or regional somatic mutation variations. For the genomic regions, we either used all of the genomic regions or sub-selected 5% genomic regions that represent the largest difference in the regression model between H3K4me1 liver and stomach mucosa. Levels for expression factors (DR, MDTC) of genes in each cluster were aggregated by 1-megabase window for all genomic regions with DR cutoff of >0.05 or selected genomic regions without the cutoffs. If a particular gene spans two 1-megabase genomic regions, we applied the aggregation of expression factor levels on the region where the gene has a greater length proportion.

4.8. Prediction of cell-of-origin by utilizing scRNA-seq data

In order to complement the chromatin feature-based COO predictions, we applied the previous random forest algorithm by substituting the chromatin features into the scRNA-seq data of human liver tissues. scRNA-seq data from a total of 20 single cell clusters (6 hepatocytes clusters, 1 cholangiocyte cluster, 3 endothelial cells clusters, 1 hepatic stellate cells cluster, 2 B cells clusters, 3 T cells clusters, 1 NK-like cells cluster, 2 intrahepatic monocyte/macrophage clusters, and 1 erythrocyte cluster) generated from previous study (MacParland et al., 2018) were used for the COO prediction, and the DR expression factor values derived from each cluster were added up based on the gene distribution in 1-megabase window (same windows as chromatin features) for all genomic regions. Eventually, from the variables of these 20 clusters sorted by 1-megabase window, we applied greedy backward elimination to figure out the most significant cluster for the regional mutation density of each sample. For our purpose, we considered the samples with positive variance explained score as relevant samples for the COO assignments. In case of predicting COO for each PLCs subtype of aggregated samples, we applied greedy backward elimination using the average DR value of clusters corresponding to each cell type and subsequently ranked the DR value features for each cell type.

4.9. Data availability

The authors declare that all data supporting the findings of this study are available within the paper and its supplementary information files.

Declarations

Author contribution statement

H. Lee and K. Ha: Conceived and designed the experiments; Analyzed and interpreted the data; Wrote the paper.

H. Kim: Conceived and designed the experiments.

M. Fujita and R. Karlić: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

S. Yang; Analyzed and interpreted the data.

R. Xue, C. Zhang, F. Bai, N. Zhang, P. Polak, Y. Hoshida and H. Nakagawa:Contributed reagents, materials, analysis tools or data.

Funding statement

This work was supported by the 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), a subsidiary agency under MSIP (No. NRF-2017R1A4A1014584, Epigenetic Regulation of Bone & Muscle Regeneration Lab).

Competing interest statement

The authors declare the following conflict of interests: H. Lee is currently working at UPPThera, Inc., but conducted the current research without any conflict of financial interests. Other authors declare no competing financial interests.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2020.e03350.

References

Asrani, S.K., Devarbhavi, H., Eaton, J., Kamath, P.S., 2019. Burden of liver diseases in the world. J. Hepatol. 70, 151–171.

Blokzijl, F., de Ligt, J., Jager, M., Sasselli, V., Roerink, S., Sasaki, N., Huch, M., Boymans, S., Kuijk, E., Prins, P., et al., 2016. Tissue-specific mutation accumulation in human adult stem cells during life. Nature 538, 260–264.

Blokzijl, F., Janssen, R., van Boxtel, R., Cuppen, E., 2018. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Med. 10, 33.

Blum, H.E., Stowring, L., Figus, A., Montgomery, C.K., Haase, A.T., Vyas, G.N., 1983. Detection of hepatitis B virus DNA in hepatocytes, bile duct epithelium, and vascular elements by in situ hybridization. Proc. Natl. Acad. Sci. U. S. A. 80, 6685–6688.

Cardinale, V., Wang, Y., Carpino, G., Cui, C.B., Gatto, M., Rossi, M., Berloco, P.B., Cantafora, A., Wauthier, E., Furth, M.E., et al., 2011. Multipotent stem/progenitor cells in human biliary tree give rise to hepatocytes, cholangiocytes, and pancreatic islets. Hepatology 54, 2159–2172.

Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., Getz, G., 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. 31, 213–219.

Consortium, E.P., 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74.

Fan, B., Malato, Y., Calvisi, D.F., Naqvi, S., Razumilava, N., Ribback, S., Gores, G.J., Dombrowski, F., Evert, M., Chen, X., et al., 2012. Cholangiocarcinomas can originate from hepatocytes in mice. J. Clin. Invest. 122, 2911–2915.

Fletcher, N.F., Humphreys, E., Jennings, E., Osburn, W., Lissauer, S., Wilson, G.K., van, I.S.C., Baumert, T.F., Balfe, P., Afford, S., et al., 2015. Hepatitis C virus infection of cholangiocarcinoma cell lines. J. Gen. Virol. 96, 1380–1388.

Fujimoto, A., Furuta, M., Totoki, Y., Tsunoda, T., Kato, M., Shiraishi, Y., Tanaka, H., Taniguchi, H., Kawakami, Y., Ueno, M., et al., 2016. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. Nat. Genet. 48, 500–509.

GBD Disease and Injury Incidence and Prevalence Collaborators, 2016. 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 388, 1545–1602.

GBD Mortality and Causes of Death Collaborators, 2016. 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 388, 1459–1544.

Guest, R.V., Boulter, L., Kendall, T.J., Minnis-Lyons, S.E., Walker, R., Wigmore, S.J., Sansom, O.J., Forbes, S.J., 2014. Cell lineage tracing reveals a biliary origin of intrahepatic cholangiocarcinoma. Can. Res. 74, 1005–1010.

Ha, K., Kim, H.G., Lee, H., 2017. Chromatin marks shape mutation landscape at early stage of cancer progression. NPJ Genom. Med. 2, 9.

Ha, K., Fujita, M., Karlić, R., Yang, S., Hoshida, Y., Polak, P., Nakagawa, H., Kim, H.-G., Lee, H.J., 2019. Somatic mutation landscape reveals differential variability of cell-of-origin for primary liver cancer. bioRxiv.

Innes, B.T., Bader, G.D., 2018. scClustViz – Single-Cell RNAseq Cluster Assessment and Visualization. F1000Research.

Jusakul, A., Cutcutache, I., Yong, C.H., Lim, J.Q., Huang, M.N., Padmanabhan, N., Nellore, V., Kongpetch, S., Ng, A.W.T., Ng, L.M., et al., 2017. Whole-genome and epigenomic landscapes of etiologically distinct subtypes of cholangiocarcinoma. Canc. Discov. 7, 1116–1135.

Kübler, K., Karlić, R., Haradhvala, N.J., Ha, K., Kim, J., Kuzman, M., Jiao, W., Gakkhar, S., Mouw, K.W., Braunstein, L.Z., et al., 2019. Tumor Mutational Landscape Is a Record of the Pre-malignant State. bioRxiv.

MacParland, S.A., Liu, J.C., Ma, X.Z., Innes, B.T., Bartczak, A.M., Gage, B.K., Manuel, J., Khuu, N., Echeverri, J., Linares, I., et al., 2018. Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. Nat. Commun. 9, 4383.

Matsumoto, T., Takai, A., Eso, Y., Kinoshita, K., Manabe, T., Seno, H., Chiba, T., Marusawa, H., 2017. Proliferating EpCAM-positive ductal cells in the inflamed liver give rise to hepatocellular carcinoma. Canc. Res. 77, 6131–6143.

Michalopoulos, G.K., Barua, L., Bowen, W.C., 2005. Transdifferentiation of rat hepatocytes into biliary cells after bile duct ligation and toxic biliary injury. Hepatology 41, 535–544.

Moeini, A., Sia, D., Bardeesy, N., Mazzaferro, V., Llovet, J.M., 2016. Molecular pathogenesis and targeted therapies for intrahepatic cholangiocarcinoma. Clin. Canc. Res. 22, 291–300.

Moeini, A., Sia, D., Zhang, Z., Camprecios, G., Stueck, A., Dong, H., Montal, R., Torrens, L., Martinez-Quetglas, I., Fiel, M.I., et al., 2017. Mixed hepatocellular cholangiocarcinoma tumors: cholangiolocellular carcinoma is a distinct molecular entity. J. Hepatol. 66, 952–961.

Monga, S.P., 2019. Updates on hepatic homeostasis and the many tiers of hepatobiliary repair. Nat. Rev. Gastroenterol. Hepatol. 16, 84–86.

Mu, X., Espanol-Suner, R., Mederacke, I., Affo, S., Manco, R., Sempoux, C., Lemaigre, F.P., Adili, A., Yuan, D., Weber, A., et al., 2015. Hepatocellular carcinoma originates from hepatocytes and not from the progenitor/biliary compartment. J. Clin. Invest. 125, 3891–3903.

Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Rynes, E., Maurano, M.T., Vierstra, J., Thomas, S., et al., 2012. BEDOPS: high-performance genomic feature operations. Bioinformatics 28, 1919–1920.

Polak, P., Lawrence, M.S., Haugen, E., Stoletzki, N., Stojanov, P., Thurman, R.E., Garraway, L.A., Mirkin, S., Getz, G., Stamatoyannopoulos, J.A., et al., 2014. Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. Nat. Biotechnol. 32, 71–75.

Polak, P., Karlic, R., Koren, A., Thurman, R., Sandstrom, R., Lawrence, M., Reynolds, A., Rynes, E., Vlahovicek, K., Stamatoyannopoulos, J.A., et al., 2015. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. Nature 518, 360–364.

Raven, A., Lu, W.Y., Man, T.Y., Ferreira-Gonzalez, S., O'Duibhir, E., Dwyer, B.J., Thomson, J.P., Meehan, R.R., Bogorad, R., Koteliansky, V., et al., 2017. Cholangiocytes act as facultative liver stem cells during impaired hepatocyte regeneration. Nature 547, 350–354.

Razumilava, N., Gores, G.J., 2014. Cholangiocarcinoma. Lancet 383, 2168–2179.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K., 2015. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43, e47.

Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al., 2015. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330.

Russell, J.O., Lu, W.Y., Okabe, H., Abrams, M., Oertel, M., Poddar, M., Singh, S., Forbes, S.J., Monga, S.P., 2019. Hepatocyte-specific beta-catenin deletion during severe liver injury provokes cholangiocytes to differentiate into hepatocytes. Hepatology 69, 742–759.

Sekiya, S., Suzuki, A., 2012. Intrahepatic cholangiocarcinoma can arise from Notch-mediated conversion of hepatocytes. J. Clin. Invest. 122, 3914–3918.

Sekiya, S., Suzuki, A., 2014. Hepatocytes, rather than cholangiocytes, can be the major source of primitive ductules in the chronically injured mouse liver. Am. J. Pathol. 184, 1468–1478.

Sia, D., Villanueva, A., Friedman, S.L., Llovet, J.M., 2017. Liver cancer cell of origin, molecular class, and effects on patient prognosis. Gastroenterology 152, 745–761.

Tummala, K.S., Brandt, M., Teijeiro, A., Grana, O., Schwabe, R.F., Perna, C., Djouder, N., 2017. Hepatocellular carcinomas originate predominantly from hepatocytes and benign lesions from hepatic progenitor cells. Cell Rep. 19, 584–600.

Vicent, S., Lieshout, R., Saborowski, A., Verstegen, M.M.A., Raggi, C., Recalcati, S., Invernizzi, P., van der Laan, L.J.W., Alvaro, D., Calvisi, D.F., et al., 2019. Experimental models to unravel the molecular pathogenesis, cell of origin and stem cell properties of cholangiocarcinoma. Liver Int. 39 (Suppl 1), 79–97.

Wang, B., Zhao, L., Fish, M., Logan, C.Y., Nusse, R., 2015. Self-renewing diploid Axin2(+) cells fuel homeostatic renewal of the liver. Nature 524, 180–185.

Wang, Z., Sheng, Y.Y., Dong, Q.Z., Qin, L.X., 2016. Hepatitis B virus and hepatitis C virus play different prognostic roles in intrahepatic cholangiocarcinoma: a meta-analysis. World J. Gastroenterol. 22, 3038–3051.

Wang, J., Dong, M., Xu, Z., Song, X., Zhang, S., Qiao, Y., Che, L., Gordan, J., Hu, K., Liu, Y., et al., 2018. Notch2 controls hepatocyte-derived cholangiocarcinoma formation in mice. Oncogene 37, 3229–3242.

Wardell, C.P., Fujita, M., Yamada, T., Simbolo, M., Fassan, M., Karlic, R., Polak, P., Kim, J., Hatanaka, Y., Maejima, K., et al., 2018. Genomic characterization of biliary tract cancers identifies driver genes and predisposing mutations. J. Hepatol. 68, 959–969.

Xue, R., Chen, L., Zhang, C., Fujita, M., Li, R., Yan, S.M., Ong, C.K., Liao, X., Gao, Q., Sasagawa, S., et al., 2019. Genomic and transcriptomic profiling of combined hepatocellular and intrahepatic cholangiocarcinoma reveals distinct molecular subtypes. Canc. Cell 35, 932–947 e938.

Yanger, K., Zong, Y., Maggs, L.R., Shapira, S.N., Maddipati, R., Aiello, N.M., Thung, S.N., Wells, R.G., Greenbaum, L.E., Stanger, B.Z., 2013. Robust cellular reprogramming occurs spontaneously during liver regeneration. Genes Dev. 27, 719–724.

Zender, S., Nickeleit, I., Wuestefeld, T., Sorensen, I., Dauch, D., Bozko, P., El-Khatib, M., Geffers, R., Bektas, H., Manns, M.P., et al., 2013. A critical role for notch signaling in the formation of cholangiocellular carcinomas. Canc. Cell 23, 784–795.

Ziogas, D.E., Kyrochristos, I.D., Glantzounis, G.K., Christodoulou, D., Felekouras, E., Roukos, D.H., 2017. Primary liver cancer genome sequencing: translational implications and challenges. Expet Rev. Gastroenterol. Hepatol. 11, 875–883.