의학박사 학위논문

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

식도암과 간암의 기원세포 다양성에 대한 연구

2020년 8월

서울대학교 대학원

의료정보학 협동과정

하 경 식

Abstract

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

Kyungsik Ha
Interdisciplinary Program of Medical Informatics
The Graduate School
Seoul National University

Primary esophageal and liver cancers display consistent increase in the global disease burden and mortality. Identification of cell-of-origins for primary cancers would be a necessity to expand options for designing relevant therapeutics and preventive medicine for these cancer types. Although multiple studies addressed possible heterogeneity of cell-of-origins for these cancers and its subtypes, an integrative research on cell-of-origin for each cancer utilizing human specimen data was poorly established. To this end, we analyzed previously published whole-genome sequencing data for pre-cancerous, cancer, and progenitor tissues along with publicly available normal tissue epigenomic features to conduct *in-silico* prediction of the cell-of-origin for primary cancer subtypes. Especially in the case of primary liver cancer, we included single cell RNA-seq data from human livers to assess correlation patterns and verified this information from cell-of-origin analysis at cell level. Our data showed that the establishment of somatic mutation landscape inferred by chromatin features occur early in the process of cancer

progression, and gastric acid reflux environmental exposure-mediated epigenetic changes, represented as gastric metaplasia, at early stage can dramatically impact on determining cell-of-origin of esophageal cancer. In addition, despite mixed histological features, the cell-of-origin for mixed hepatocellular carcinoma/intrahepatic cholangiocarcinoma subtype was predominantly predicted to be hepatocytic origin. Furthermore, 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. Furthermore, a rare proportion of hepatocellular carcinomas were predicted as a non-hepatocytic cell-of-origin, which also demonstrated a high expression level of epithelial cell marker specific marker, *EPCAM*. Collectively, these results provide novel insights on the heterogeneous nature and potential contributors of cell-of-origin for primary cancers.

---

keywords: somatic mutation, epigenetic features, cell-of-origin, machine learning, esophageal cancer, primary liver cancer
Student Number: 2011-23816

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLE

Introduction

1. Studies on the relationship between somatic mutation landscapes and epigenetic features in cancer genomics

Recent advances in cancer genomics have so far revealed numerous somatic mutation landscapes for various cancer types, leading to a number of key findings. Identification of new driver gene mutations, deciphering clonal evolution structure, and profiling tumor heterogeneity within and among different patients through examination of mutations, mainly at the gene level (Alexandrov et al., 2013; Hodgkinson et al., 2012; Kan et al., 2010; Kandoth et al., 2013; Lawrence et al., 2013; Martincorena and Campbell, 2015; Schaefer and Serrano, 2016), have successfully addressed the genes contributing to cancer progression and identified novel therapeutic targets. Beyond these gene-focused approaches, systematic analyses of mechanisms that could explain genomic regional variations in mutation rates across various cancer types could significantly extend our understanding about common contributors to the establishment of mutation landscapes before and during cancer progression. To this end, a number of studies have examined relationships between regional mutation frequencies across the genome and some types of features, including gene expression level and DNA sequence context. While these factors are being investigated intensively, the studies on the relationship between somatic mutation profiles and other features such as epigenetic marks has been thought to be less significant in cancer genomics so far. Epigenetic marks are heritable elements that can affect the phenotype without altering the DNA sequence. Although DNA methylation was considered to be the only epigenetic marker that can be inherited in the traditional criteria, recent definitions of epigenetic marks include histone post-translational modifications (methylation, acetylation, phosphorylation, ubiquitination,

etc.) and histone variants (H3, H4, H2A, H2B) to deal with the complexity of living organisms. Since these marks contribute to the regulation of gene expression and DNA replication by affecting the chromatin environment such as high order chromatin structures, nucleosome occupancy, and hypomethylated blocks in the nucleus, both epigenetic marks and chromatin changes are crucial to identify cell characteristics and their cancer progression. In line with this, recent studies have also investigated the relatedness between somatic mutation profiles and these epigenetic features, including histone post-translational modifications and open chromatin marks such as DNase1-seq profiles (Liu et al., 2013; Polak et al., 2015; Polak et al., 2014; Schuster-Bockler and Lehner, 2012; Stamatoyannopoulos et al., 2009; Supek and Lehner, 2015; Thurman et al., 2012; Woo and Li, 2012). In consequence, it was confirmed that these epigenetic features display high correlation with regional mutation rates.

2. Predicting cell-of-origin (COO) of cancer at subtype and individual sample level

Among the various studies applying an intimate relationship between somatic mutation profiles and epigenetic features, one of the successful cases was the development of an algorithm for predicting cell-of-origin (COO) of cancer (Polak et al., 2015). Although this study was designed to predict the mutation rate of each cancer cell type from the epigenetic features of normal cells at the 1 megabase-level, the fundamental goal of the study was to trace the cell-of-origin of cancer from the importance ranking of epigenetic features that contribute to predict the mutation rate using random forest regression. Until now, a total of eight cancer types was investigated by aggregated the somatic mutation data, and the

cell-of-origin of each cancer type was correctly predicted except for the lung cancer type that has no associated epigenetic data (Polak et al., 2015). However, despite the successful development of the COOs prediction algorithm for each cancer type, no studies were yet performed to cancer subtype-specific or individual sample-specific predictions that could have clinical implications associated with prevention strategies. COOs classification of cancer at subtype and individual sample level is clinically important, because it has been confirmed that some cancer type could have multiple COOs from different cells/tissues through animal experiments and cell line studies. In other some cancer type cases, cancerous tissues can be initiated from fully differentiated cells or from progenitor cells. Thus, it is crucial to distinguish COOs depending on subtype or individual sample to establish the early-stage diagnosis and possibly the treatment selection for each case. This distinction is also essential to understand the biological mechanism of cancer, since cancer progression could differ depending on the origin cells of the cancer.

Here, we performed a computational approach to dissect out the putative COOs on each cancer subtype and interrogated possible individual tumor-level heterogeneity in COOs. For this, we analyzed a total of whole genome sequencing data from barrett's esophagus (BE), esophageal adenocarcinoma (EAC), esophageal squamous cell carcinoma (ESCC), primary liver cancer (PLC), extrahepatic biliary tract cholangiocarcinoma (BTCA), and hepatic progenitor cells (HPC) for assessing the possibility as a common COO for PLCs, along with 423 of chromatin features at the epigenome-level. 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. 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 each cancer type, but also uncovering the differential contribution of each COO in different subtypes of each cancer type.

Materials and Methods

Data for esophageal cancer. For the purposes of our project, we used somatic mutation data from BE, EAC, and ESCC tissues. Data use were authorized from ICGC (http://icgc.org) and BGI (http://www.genomics.cn/) before use. A total of 23 pairs of Barrett's esophagus and matching esophageal adenocarcinoma genomics data (Ross-Innes et al., 2015) were authorized from ICGC and genome data of 14 ESCC samples (Zhang et al., 2015) were acquired from BGI. These data sets were subsequently analyzed following the standard GATK pipeline (https://www.broadinstitute.org/gatk/) and somatic variants were called using the MuTect algorithm (Cibulskis et al., 2013)(https://www.broadinstitute.org/cancer/cga/mutect).

Data for primary liver cancer. 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 and as an independent cohort for predicting the COO of ICCs.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.

Epigenomic data and data processing. A total of 423 epigenomics and chromatin data were from the NIH Roadmap Epigenomics Mapping Consortium (Roadmap Epigenomics et al., 2015) and ENCODE (Consortium, 2012). NIH Roadmap Epigenomics data were accessible from the NCBI GEO series GSE18927, referring to the University of Washington Human Reference Epigenome Mapping Project.

To calculate the regional mutation density and mean signal of chromatin features, all autosomes were split in 1-Mbp regions followed by filtering out regions containing centromeres, telomeres and low quality unique mappable base pairs. To determine regional mutation density and histone modification profiles, we counted the total number of somatic mutations or ChIP-seq reads per each 1 megabase region. For analyzing the DNase I hypersensitivity and Repli-seq data, scores of DNase I peaks and replication were calculated per each 1 megabase region. For somatic mutations, ChIP-seq data and DNase I hypersensitivity data, BEDOPS (Neph et al., 2012) was employed to calculate the frequency and scores per each 1Mbp region.

Principal coordinate analysis. Principal coordinate analysis was

used to represent differences in mutation frequency distribution among the individual samples. A dissimilarity matrix was built using 1 – Pearson correlation coefficient across all samples. Each sample location was assigned in a two-dimensional space using this matrix.

Feature selection based on random forest regression. A random forest regression-based feature selection algorithm was performed as described (Polak et al., 2015) with modifications. Briefly, the training set for each tree was constructed, followed by using out-of-bag data to estimate the mean squared error. Thus, there was no need to perform additional tests for error evaluation. Out-of-bag data were also used to estimate the importance of each variable. In each out-of-bag case, the values corresponding to each variable were randomly permuted, then tested to each tree. Subtracting the score of the mean squared error between the untouched out-of-bag data cases and the variable-m-permuted cases, the raw importance score of variable m was measured. By calculating the average score of variable m in the entire tree, the rank of importance for each variable was determined. A total of 1,000 random forest trees were employed to predict mutation density using a total of 423 chromatin features. Every random forest model was repeated 1,000 times.

After the random forest algorithm step, greedy backward elimination was performed to select the top 20 chromatin variables. Subsequent removal of the lowest rank variable was done to calculate the variance explained value measurements for each variable. To conduct feature selection on all of the samples corresponding to the particular pre-cancerous tissues or cancer types, mutation density was calculated by adding samples in each case. However, the subgrouping of samples was employed for specific analyses of the esophageal cancer type. To perform feature selection classified by differential

dysplasia states, samples were divided into 3 groups: 17 samples of no dysplasia, 3 samples of low-grade dysplasia and 2 samples of high-grade dysplasia. In the case of feature selection after subgrouping for distinct and common mutations, all mutations in paired-samples of BE and EAC were divided into 3 different groups: Barrett's only, EAC only, and common mutations.

Analysis of mutation frequency variance explained by chromatin features for esophageal cancer. To examine the effect of a particular cell-type specific chromatin context on explaining regional variability of mutation density across the genome, chromatin features were subgrouped based on the feature selection algorithm. To study the differences in variance explained values among distinct cell types, 9 groups were categorized. Each group included 5 chromatin markers common among the groups: H3K27me3, H3K36me3, H3K4me1, H3K4me3 and H3K9me3. Random selection of 6 chromatin features were either from all of the 423 features or 417 features (excluding stomach mucosa chromatin features). Random selection of chromatin features was repeated 1,000 times, then the average variance explained values and permutation distributions were obtained.

Prediction of regional mutation frequencies in 1-megabase genomic regions with differential chromatin levels for esophageal cancer. To select 1-megabase genomic regions with differential H3K4me1 levels, we calculated residual values derived from a linear regression model between the H3K4me1 level of stomach mucosa and that of esophagus tissue. To represent regions harboring differential H3K4me1 levels along with increased mutation accumulation rates after gastric metaplasia, a total of 92 regions were chosen based on the two criteria: (1) displaying top 5% in term of

the residual values, (2) showing higher H3K4me1 levels in esophagus than stomach mucosa. Subsequently, we built two separate regression models, and then applied the model to predict the regional mutation frequencies for the 92 regions. One regression model was between observed mutation frequencies in BE with no dysplasia and H3K4me1 level of stomach mucosa, and the other one was between observed mutation frequencies in BE with no dysplasia and H3K4me1 level of esophagus.

Prediction of cell-of-origin for primary liver cancer by grouping of chromatin features. To predict cell-of-origin 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. 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.

Signature analysis of mutational processes for primary liver cancer. Nonnegative matrix factorization (NMF) algorithm was employed to investigate mutation signatures as described in previous study (Blokzijl et al., 2018). This methodology was utilized by

factoring out frequency matrix of 96-trinucleotide mutation contexts from HCC, Mixed, ICC, BTCA-SG and HPC samples.

Gene expression analysis for primary liver cancer. RNA-Seq experiments of HCC samples were performed previously (Fujimoto et al., 2016), and the data had 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.

Prediction of cell-of-origin by utilizing scRNA-seq data for primary liver cancer. 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.

Code availability. Our core analysis code utilizing the random forest feature selection algorithm will be available on GitHub (code name: Random_forest_Ha_mutation_epi).

Results and Discussions

1. Somatic mutation landscape reveals differential variability of cell-of-origin for esophageal cancer

Precancerous tissues and matching cancers display similar regional mutation frequency profile. We performed principal coordinate analysis (PCOA) to test whether the average mutation rate differences reported previously (Ross-Innes et al., 2015) reflected in the level of 1 megabase window regional mutation frequencies. Individual BE tissues formed clusters with the EAC tissues separate from the ESCC tissues, suggesting that the matching of cancer progression history might serve as a stronger factor than the cell-of-origin context itself (Figure 1). These result shows similarity in regional variation in mutation frequencies of precancerous tissues and matching cancer types, indicating that the effect of cell-of-origin context might be cancer-type dependent.

Epigenetic shifts caused by metaplasia, driven by acid reflux, explains the establishment of the somatic mutation landscape for both BE and EAC. Cell type shift, represented as gastric metaplasia, is one of the main hallmarks in the development of BE (Hayakawa et al., 2016). Thus, one could assume that the critical time point for the establishment of the mutation landscape for BE could be either before or during the course of cell type shift, or after its completion. Chromatin feature selection analysis of the mutation landscape of BE and EAC tissues confirmed that high-ranked chromatin features were derived from the stomach tissue type (Figure S1) for both tissues, without any significant esophageal chromatin features. Simple correlation between regional mutation frequency and histone modification marks from stomach and esophagus tissues revealed marginal differences between BE and EAC tissues (Figure

S2a, b), and this pattern was also consistent with the correlation to stomach tissue DNase I hypersensitivity profile (Figure S3a). Moreover, six features covering all stomach chromatin features subjected to the feature selection analysis solely explained over 80% of the regional mutation variance for both BE and EAC tissues, which is unlikely to be non-random (p value < 2.2e-16) (Figure S4). These results imply that the major time point of mutation landscape establishment for BE is most likely to be after the cell type shift into stomach mucosa-like cells. Chromatin feature selections on subgroups of somatic mutations for BE and EAC based on overlap and uniqueness of the mutations shared common top-ranked stomach chromatin features (Figure 2a). In addition, chromatin feature selection on sample subgroups with respect to dysplasia grade revealed that the top features all originated from stomach tissue (Figure S5) and the variance explained level for all of the dysplasia-based subgroups using six stomach tissue chromatin features were similar to the variance explained level using all 423 chromatin features (Figure 2b). This finding was consistent with the high correlation to stomach tissue DNase I hypersensitivity profile (Figure S3b). Next, we sought to further determine whether the contribution of stomach mucosa chromatin features were indeed more crucial than esophagus chromatin features for shaping the mutation landscape of BE through an independent type of analysis. For this, H3K4me1 chromatin feature was used since this single feature explains most of the variance in mutation frequency of BE. Ninety-two 1-megabase regions displaying differential H3K4me1 levels were selected (methods) based on the speculation that these regions would likely to represent accelerated mutation accumulations through epigenetic changes during gastric metaplasia. Subsequently, we predicted mutation frequencies in the 92 regions by linear regression-based modeling using H3K4me1 level of

either stomach mucosa or esophagus tissue (methods). Comparing the observed and predicted mutation frequencies in the 92 regions revealed that the mutation frequencies predicted by H3K4me1 of stomach mucosa was similar to the observed regional mutation frequencies, but the mutation frequencies predicted by H3K4me1 of esophagus tissue was significantly different from the other two groups (Figure S6a). Moreover, regions with larger differences in H3K4me1 level overall display higher accuracy of mutation frequency predicted by using H3K4me1 level of stomach mucosa (Figure S6b). These result further implicate that the chromatin features from stomach mucosa provide major contribution for establishing the mutation landscape of BE, as opposed to the chromatin features of esophagus tissue, a cell-of-origin for BE. From all of these results, we infer an early time point for establishment of the mutation landscape for EAC, even prior to the occurrence of dysplasia for BE, but most likely after epigenetic changes due to gastric metaplasia.

Cell-of-origin of major chromatin features associated with mutation landscape establishment for BE, EAC, and ESCC are different. To ensure that the chromatin features shaping the mutation landscape of BE and EAC were not common to any esophageal cancer type, we analyzed the genome of ESCC, another cancer type derived from the esophageal squamous epithelium without any precancerous stages with cell type shift. Although the regional mutation frequency of ESCC correlated with histone modification marks from stomach and esophagus tissues in a similar manner (Figure S2c), chromatin feature selection revealed a subset of squamous cell type and esophagus chromatin features that were significant and distinct from BE and EAC (Figure S7). Moreover, measuring the level of variance explained values per tissue or cell

type categories showed stomach chromatin features to be the strongest ones for BE and EAC, reaching higher than 90% of the variance level explained by the 423 total chromatin features, whereas esophageal chromatin features were dominant for ESCC (Figure 3). Notably, the variance explained values for each category displayed non-significant relationship with simple correlations between the chromatin marks from different tissue or cell types (BE rs = 0.36, EAC rs = 0.36, ESCC rs = 0.18). These results imply a distinct process of mutation landscape establishment for these cancer types that varies depending on the presence of precancerous tissues with cell-type shifts.

Discussion. One thing to note is that our results display non-universal chromatin features identified as significant in different cancer types. The reason for these differences in the extent of variance explaining values for any distinct chromatin feature could be complex, and the reason might be due to the tissue type-dependent differences in the mechanisms of epigenetic regulation plus the differences in major contributing chromatin features serving as either euchromatin or heterochromatin marks. One mechanistic approach to assess the extent of chromatin features contributing to mutation landscape is using CRISPR-Cas9 system to incorporate mutations on chromatin enzymes leading to global epigenetic changes, and then inducing somatic mutations using various types of mutagens to examine the effect of different epigenetic features on shaping mutation landscape, which could be one of the strong candidates for any follow-up research.

Finally, analyses results from BE and EAC raise the possibility that epigenetic changes due to environmental insults, represented as a cell type shift, could serve as a primary role for establishing the mutation

landscape of at early stage of cancer progression. Although there are possibilities that esophagus tissue chromatin features could still be involved in shaping the mutation landscape of BE in a minor manner, our analyses demonstrated that the stomach tissue chromatin features serve as a key factor shaping regional variations in somatic mutation frequency of BE.

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

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., 2015; Polak et al., 2014) 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 ($P < 2.2e-16$, Mann-Whitney U-test between the first and second rank features of each PLC subtype; Figure 4a). 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 S8a, 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 S9a). 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 S9b). 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 4b). Collectively, these results demonstrate that COO-associated chromatin features can delineate the relationships with the mutation landscape of PLCs and BTCAs.

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 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 S10a). 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 S10b).

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 4a) 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 S10c). 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.

3. Functional study for primary liver cancer

Individual Sample-level Cell-of-origin Predictions for primary liver cancer. 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 5a). 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 S11a), 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 S11b), 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 S12). For ICCs, only 5 out of 56 samples displayed positive variance explained scores, further implicating 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 S13). 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 5b), 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 5c). Clustering analysis confirmed that HCCs with non-hepatocytic predicted COO were enriched in a cluster that expressed more *EPCAM* and *KRT19* (Figure 5d). Gene set enrichment analysis showed that molecular pathways associated with bile acid synthesis and xenobiotic degradation were down-regulated in HCCs with non-hepatocytic predicted COO (Figure S14). This result indicates that the functional similarity to hepatocytes is being less observed in HCCs with non-hepatocytic predicted COO. Furthermore, we identified the connection network between these two molecular pathways related to the liver function by employing the pathway intersection function in Bio-Entity Explorer (Jung et al., 2020). Then, it was confirmed that Aldo-keto reductase family 1 (AKR1) involved in steroid metabolism was a common enzyme between bile acid synthesis and xenobiotic degradation (Figure S15). 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.

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 6a). 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.

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 S16a-c). Next, we assessed whether the proportion of signature D correlates with COO assignment for PLCs. As demonstrated in Figure 6b, 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 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 S16d). 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.

Discussion. In this study, 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., 2015; Polak et al., 2014). 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.

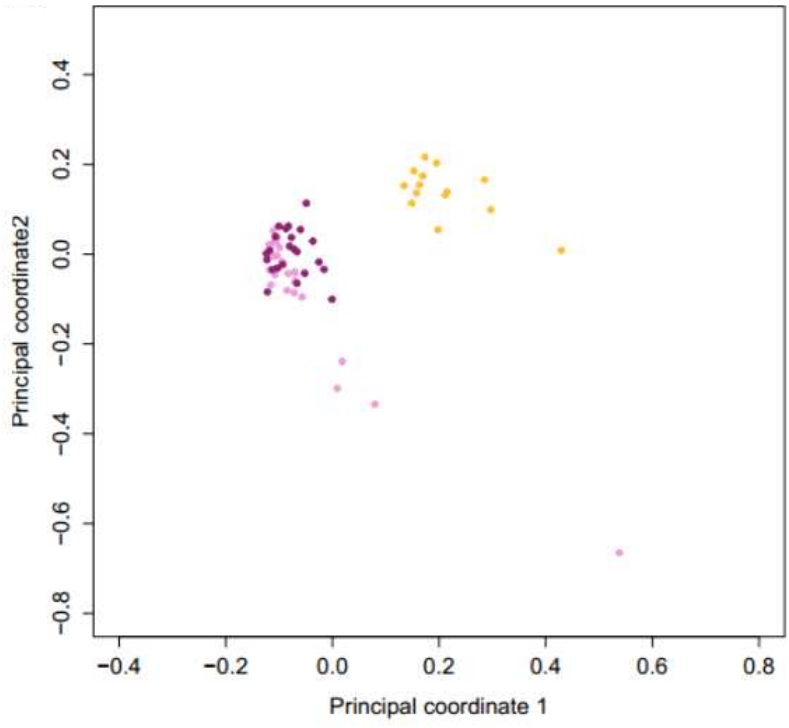
Several publications provided pieces of evidence on the

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 vitroand in vivodata 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.



	Cell Type	n
	Esophageal adenocarcinoma	23
	Barrett's esophagus	23
	ESCC	14

Figure 1. Principal coordinate analysis (PCOA) of individual cancer samples. Barrett's esophagus, esophageal adenocarcinoma, and ESCC.

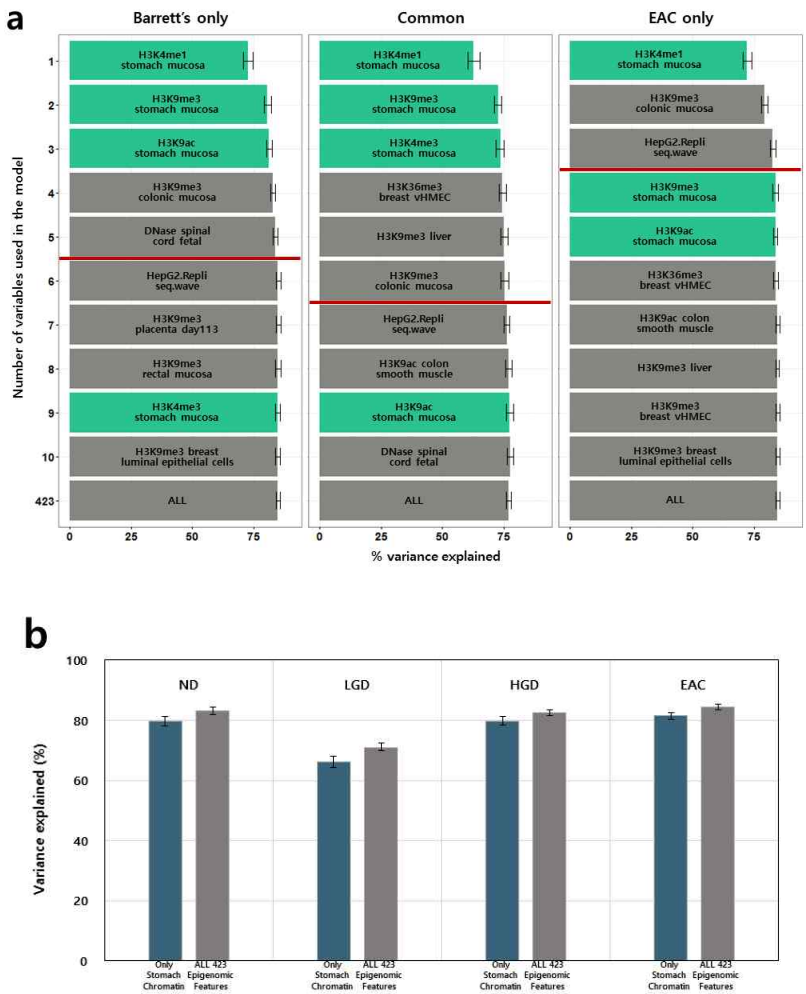


Figure 2. Regional mutation frequency landscape of Barrett's esophagus and matching esophageal adenocarcinoma are affected by cell-type-shift-associated epigenetic changes. (a) Chromatin feature selection based on the commonality of mutations in paired samples of Barrett's esophagus and esophageal adenocarcinoma. Barrett's only: mutations observed only in the Barrett's esophagus genome, Common: mutations observed in common for both Barrett's esophagus and esophageal Adenocarcinoma genomes, EAC only: mutations observed solely in the

esophageal adenocarcinoma genome. (b) Bar graph representing average variance explained scores using either stomach chromatin features (navy) or all 423 epigenomic features (gray). ND: no dysplasia, LGD: low-grade dysplasia, HGD: high-grade dysplasia, EAC: esophageal adenocarcinoma. Error bars demonstrate minimum and maximum values derived from 1,000 repeated simulations.

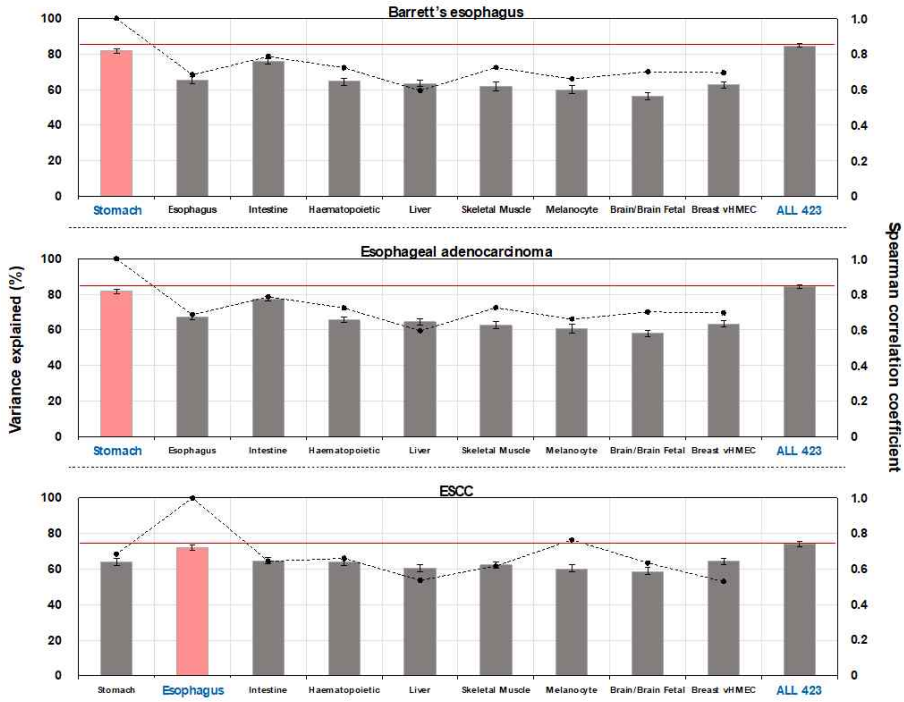


Figure 3. Regional mutation frequency landscape of esophageal squamous cell carcinoma demonstrates the uniqueness of significant chromatin features associated with the Barrett's esophagus and esophageal adenocarcinoma genomes. Average variance explained scores for pre-cancerous or matching cancer genomes were separately calculated using the tissue or cell type-based subgroup-classified chromatin features. The pink panel represents subgroups with the highest variance explained score for each cell type. The red line indicates the variance explained score when using all 423 epigenomic features. Dots represent the Spearman's rank correlations (r) of chromatin features between the highest variance explained-scored subgroup and the remaining subgroups. Error bars demonstrate minimum and maximum values derived from 1,000 repeated simulations.

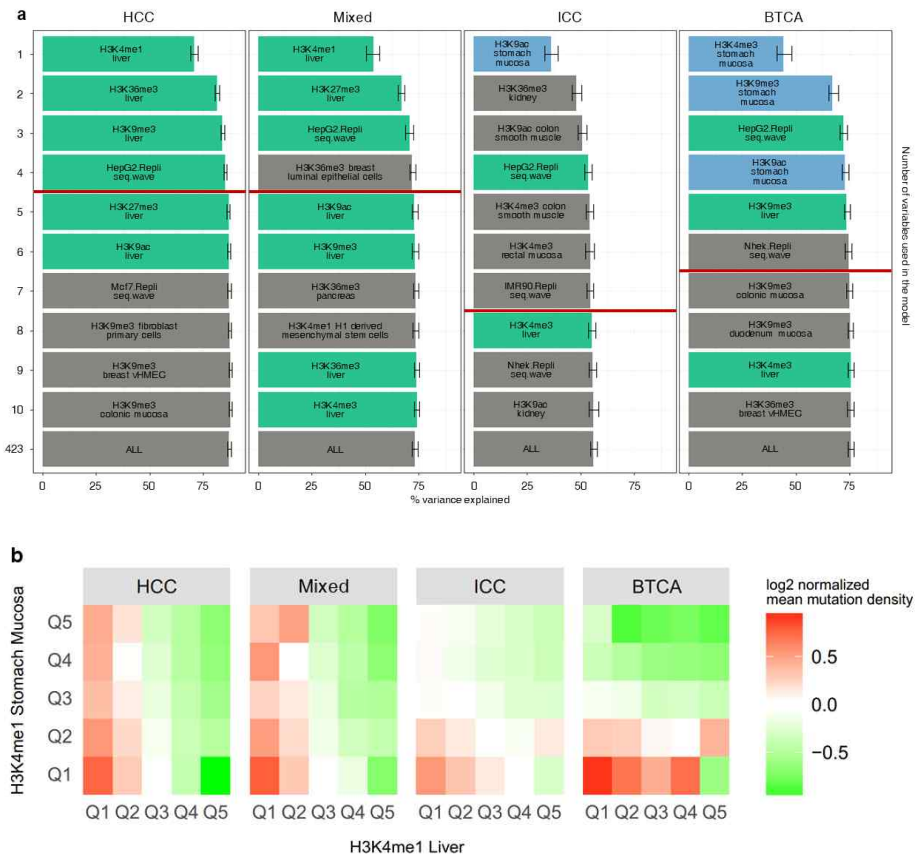


Figure 4. Cell-of-origin chromatin features delineating relations with the regional mutation frequency of HCCs, Mixed, ICCs and BTCAs. (a) Random forest regression-based chromatin feature selection using aggregated somatic mutation frequency data from HCC, Mixed, ICC and BTCA-SG samples. The rank of each chromatin feature wasdetermined by importance values. Barlength represents the variance explained scores, and theerror 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.

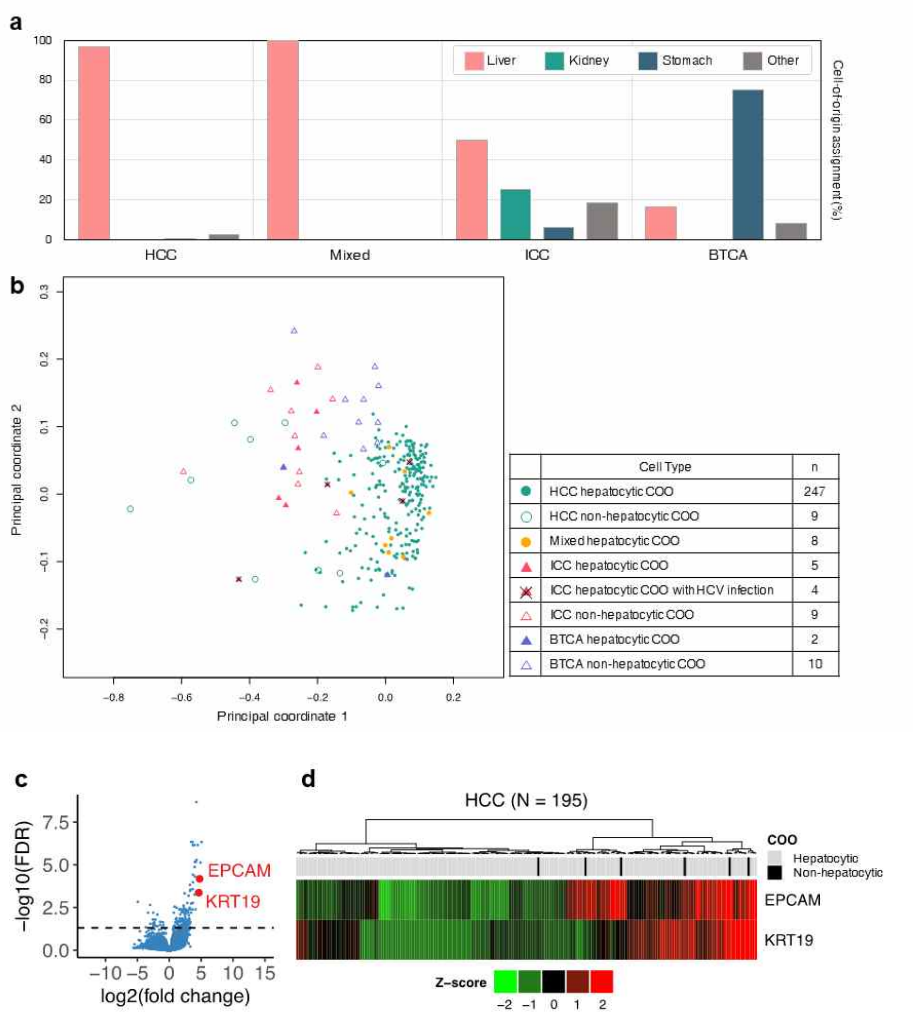


Figure 5. Analysis of COOs forindividual cancer samples. (a) Prediction of COO via grouping of chromatin features for each normal tissue type. The bar graph depictsthe percentage of samples with respect to the assigned COO by liver tissue chromatin features (pink), kidney tissue chromatin features (green), stomach tissue chromatin features (navy) or the rest (gray). (b) Principal coordinate analysisof 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.

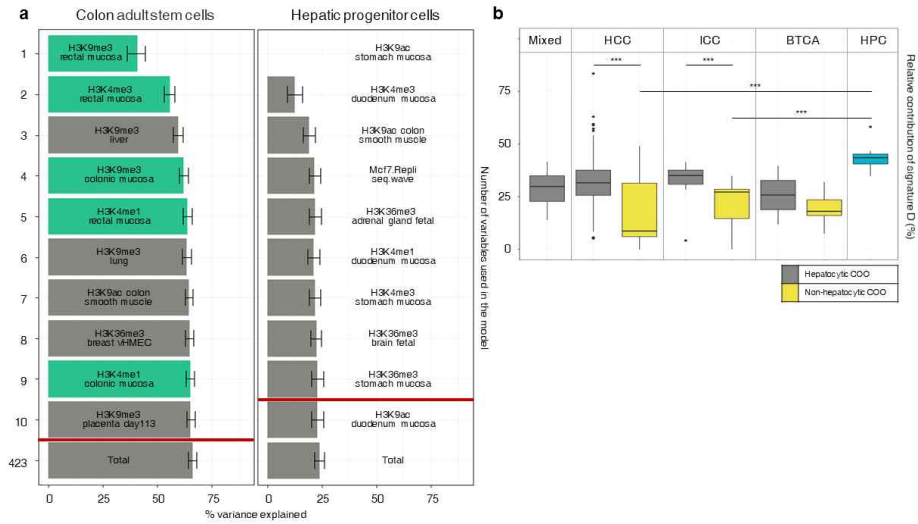


Figure 6. Hepatic progenitor cells displaydistinct mutation landscape and mutational signature processes compared to the genomes of PLCs. (a) Chromatin feature selection in relation to the regional mutation frequency of colon adult stem cells and hepatic progenitor cells. The chromatin features related 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 wascalculated by using aMann-Whitney U-test (∗∗∗, $P < 0.05$). BTCAs were excluded from the statistical analysis because only two samples were predicted as hepatocytic COO.

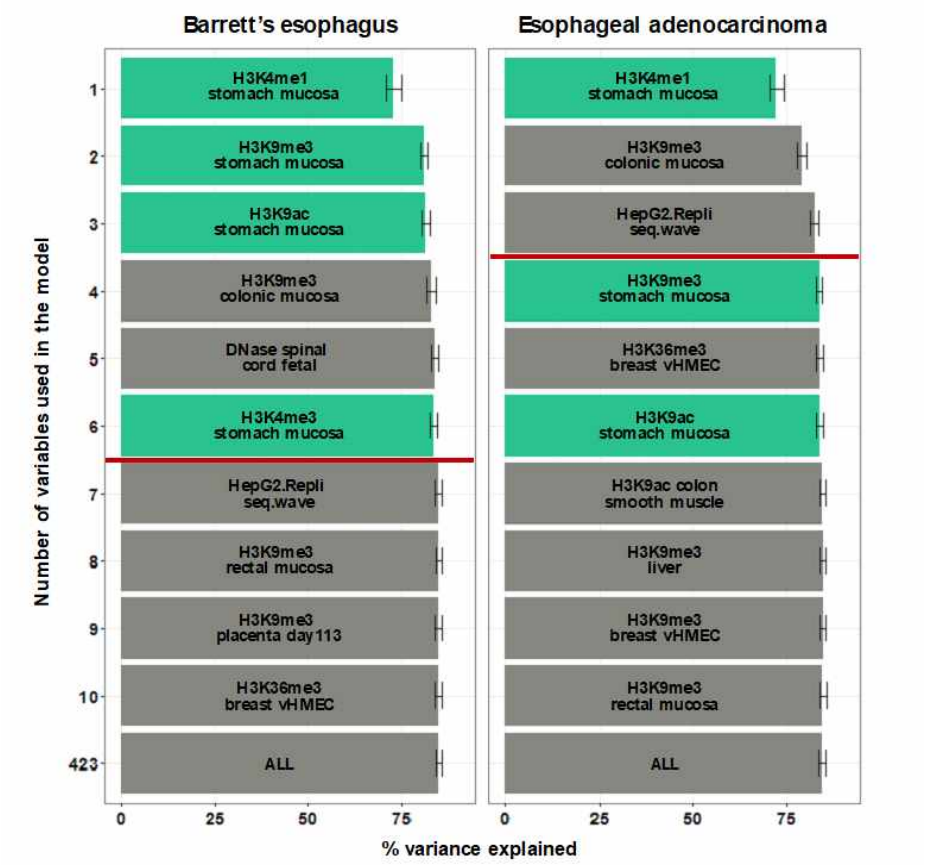
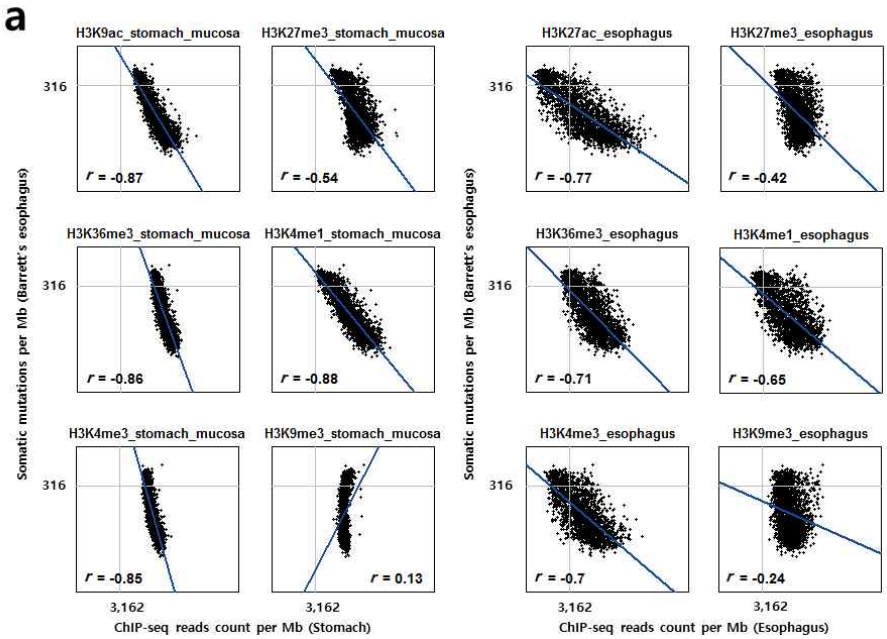
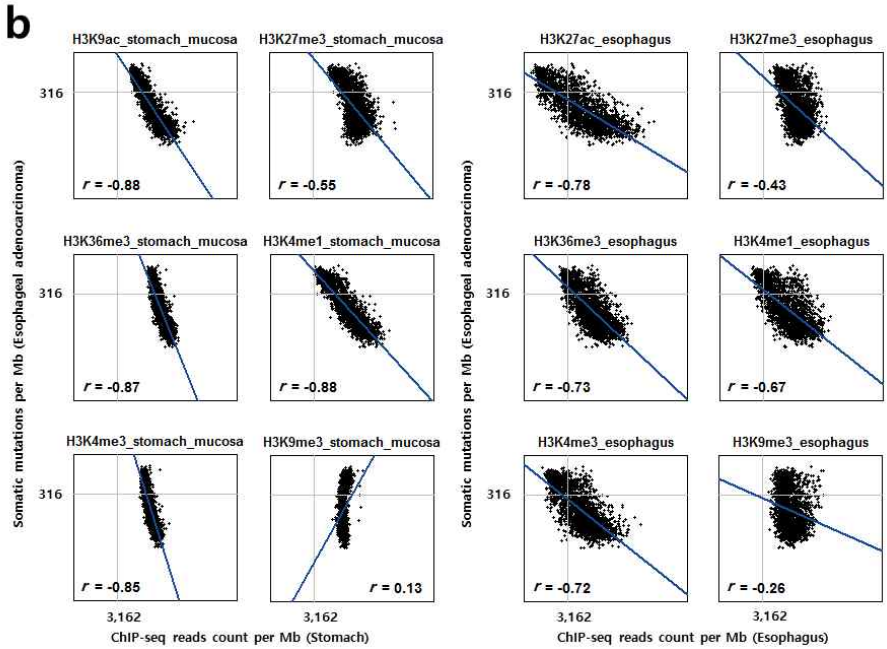


Figure S1. Chromatin feature selection in relation to the regional mutation frequency of Barrett's esophagus and esophageal adenocarcinoma. Chromatin features of the stomach mucosa are green-colored.


a
H3K9ac_stomach_mucosa
H3K27me3_stomach_mucosa
H3K27ac_esophagus
H3K27me3_esophagus
H3K36me3_stomach_mucosa
H3K4me1_stomach_mucosa
H3K36me3_esophagus
H3K4me1_esophagus
H3K4me3_stomach_mucosa
H3K9me3_stomach_mucosa
H3K4me3_esophagus
H3K9me3_esophagus
r = -0.87
r = -0.54
r = -0.77
r = -0.42
r = -0.86
r = -0.88
r = -0.71
r = -0.65
r = -0.85
r = 0.13
r = -0.7
r = -0.24
316
3,162
Somatic mutations per Mb (Barrett's esophagus)
ChIP-seq reads count per Mb (Stomach)
ChIP-seq reads count per Mb (Esophagus)


b
H3K9ac_stomach_mucosa
H3K27me3_stomach_mucosa
H3K27ac_esophagus
H3K27me3_esophagus
H3K36me3_stomach_mucosa
H3K4me1_stomach_mucosa
H3K36me3_esophagus
H3K4me1_esophagus
H3K4me3_stomach_mucosa
H3K9me3_stomach_mucosa
H3K4me3_esophagus
H3K9me3_esophagus
r = -0.88
r = -0.55
r = -0.78
r = -0.43
r = -0.87
r = -0.88
r = -0.73
r = -0.67
r = -0.85
r = 0.13
r = -0.72
r = -0.26
316
3,162
Somatic mutations per Mb (Esophageal adenocarcinoma)
ChIP-seq reads count per Mb (Stomach)
ChIP-seq reads count per Mb (Esophagus)

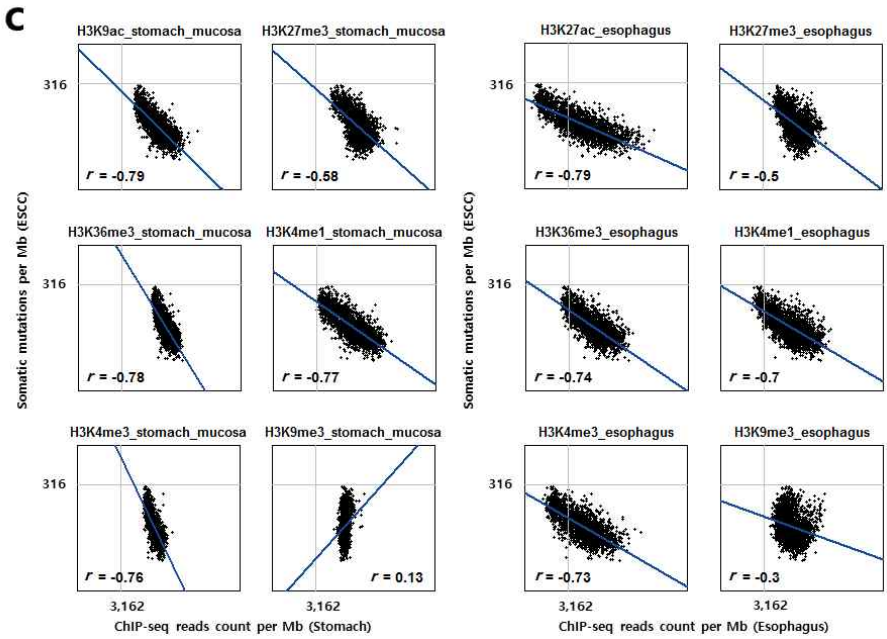


Figure S2. Correlation plots between regional mutation density and cell-type matching chromatin features. (a) Mutation density of Barrett's esophagus versus stomach mucosa or esophagus chromatin features. (b) Mutation density of esophageal adenocarcinoma versus stomach mucosa or esophagus chromatin features. (c) Mutation density of ESCC versus stomach mucosa or esophagus chromatin features.

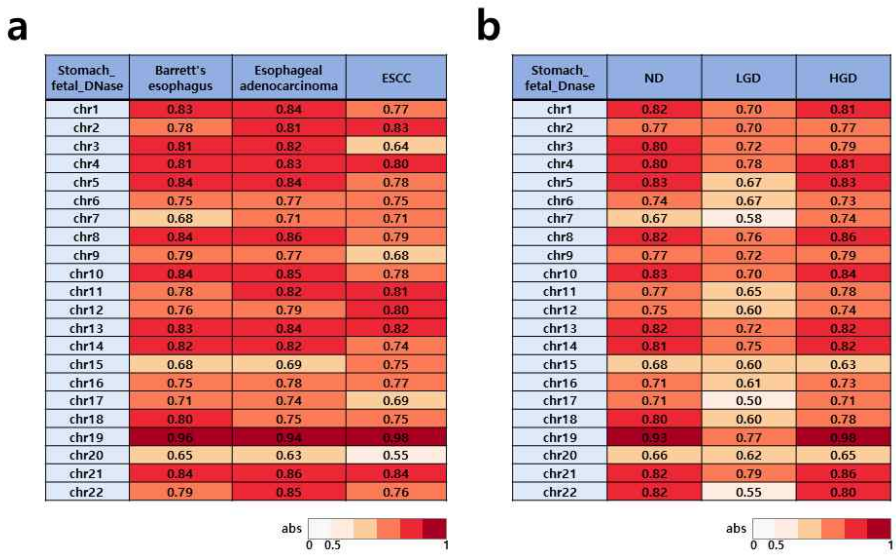


Figure S3. Spearman's rank correlation (r) between regional mutation density and chromatin accessibility index across the different chromosomes. (a) Barrett's esophagus, esophageal adenocarcinoma and ESCC. (b) Subgroups of Barrett's esophagus classified by dysplasia states.

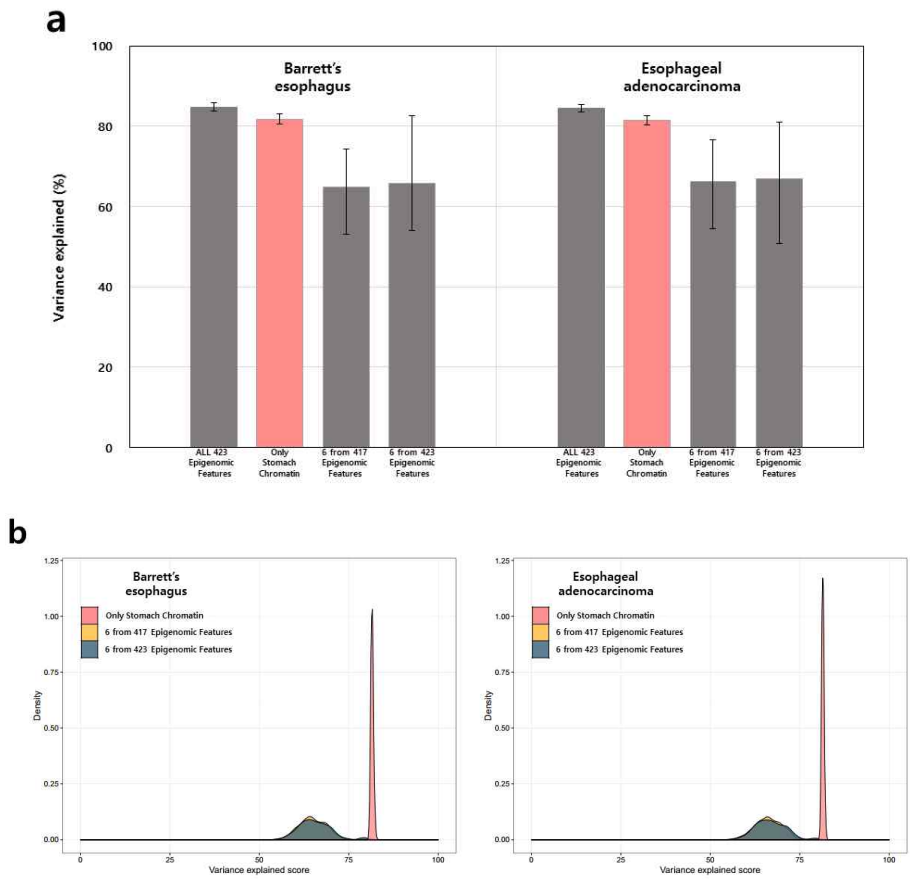


Figure S4. Comparison of variance explained scores using either stomach chromatin features or groups of randomly selected chromatin features. Stomach chromatin group represents a total of 6 chromatin features from stomach tissue. A total of 417 and 423 chromatin groups displayed 6 randomly selected chromatin features from either 417 or 423 features. The difference between 417 and 423 features was the presence or absence of stomach chromatin features. (a) Average variance explained scores using 3 different chromatin groups or all of the 423 features. Error bars demonstrate minimum and maximum values derived from 1,000 repeated simulations. (b) Distribution of variance explained scores for the group of 6 randomly selected chromatin features from either 417 or 423 chromatin features with 1,000 permutations. Pink-colored distributions represent average variance explained score of stomach chromatin features.

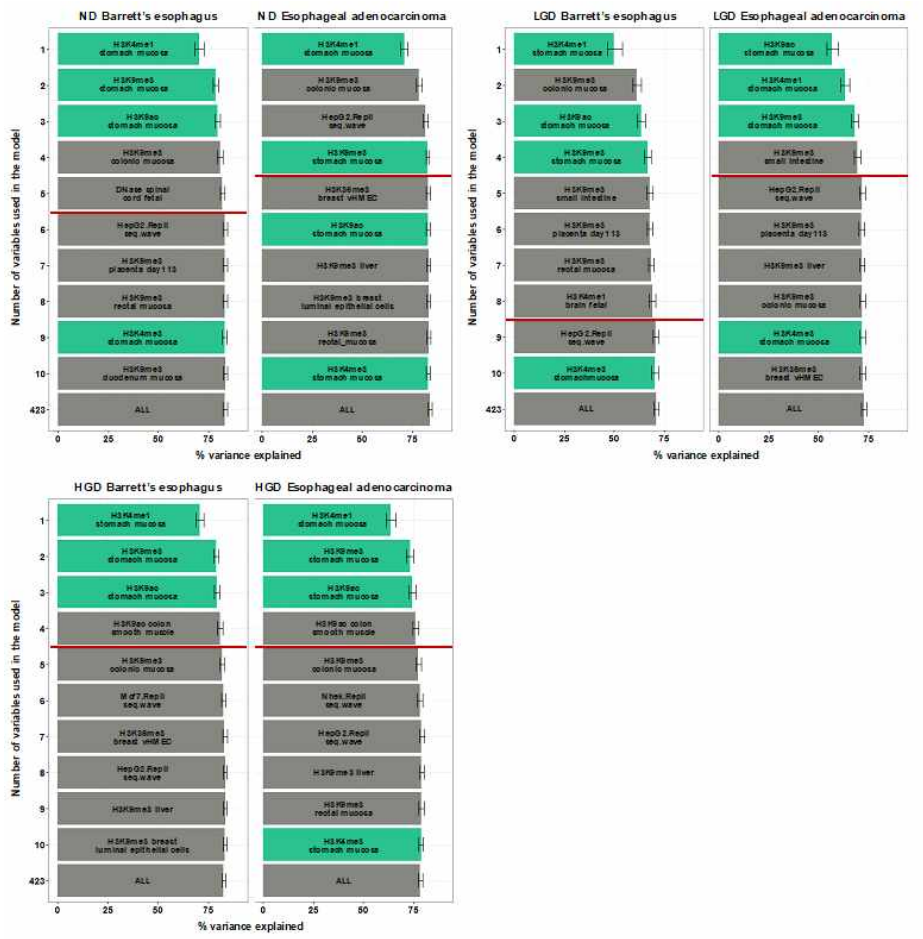


Figure S5. Feature Selection in Barrett's esophagus and esophageal adenocarcinoma classified by dysplasia status.

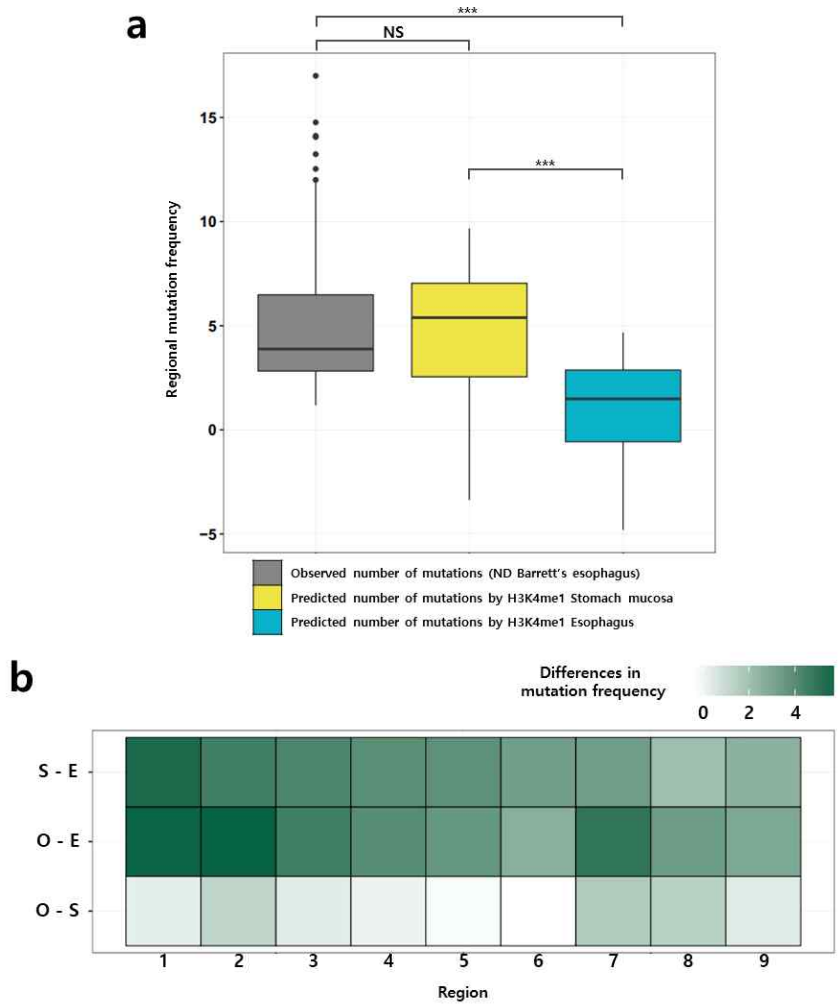


Figure S6. Comparison of observed and predicted mutation frequencies in 1 megabase genomic regions with differential chromatin level. (a) Boxplot for all 1 megabase genomic regions displaying differential chromatin level (n = 92). Statistical significance was calculated by using Krushal-Wallis one-way ANOVA followed by Dunn's test (∗∗∗, P < 0.001 ; NS, not significant). (b) Heatmap of differences in mutation frequency for the 1 megabase regions with differential chromatin level (n=92).

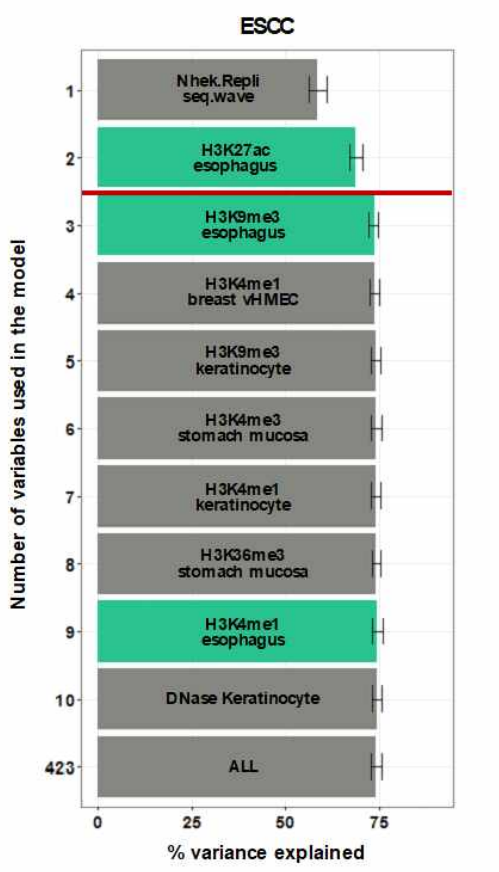


Figure S7. Chromatin feature selection in relation to the regional mutation frequency of ESCC samples. Chromatin features of the esophagus are green-colored.

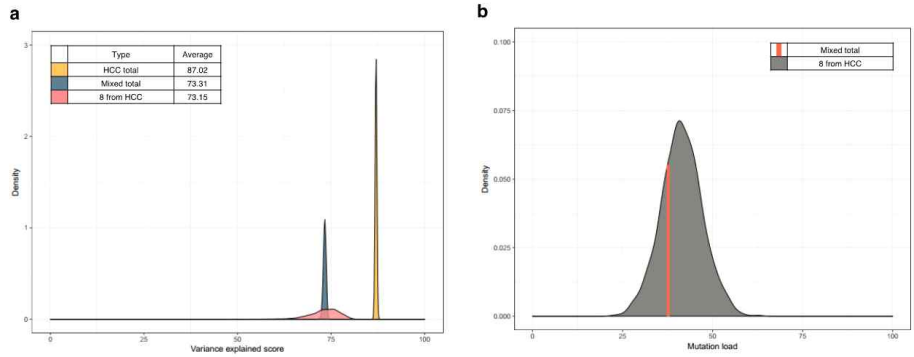


Figure S8. Difference in variance explained scores between the HCC and MIXED type is related to the total number of samples and the aggregated mutation load. (a) Distribution of variance explained scores using either all samples or 8 randomly selected samples in 1,000 repeated simulations. Distributions 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.

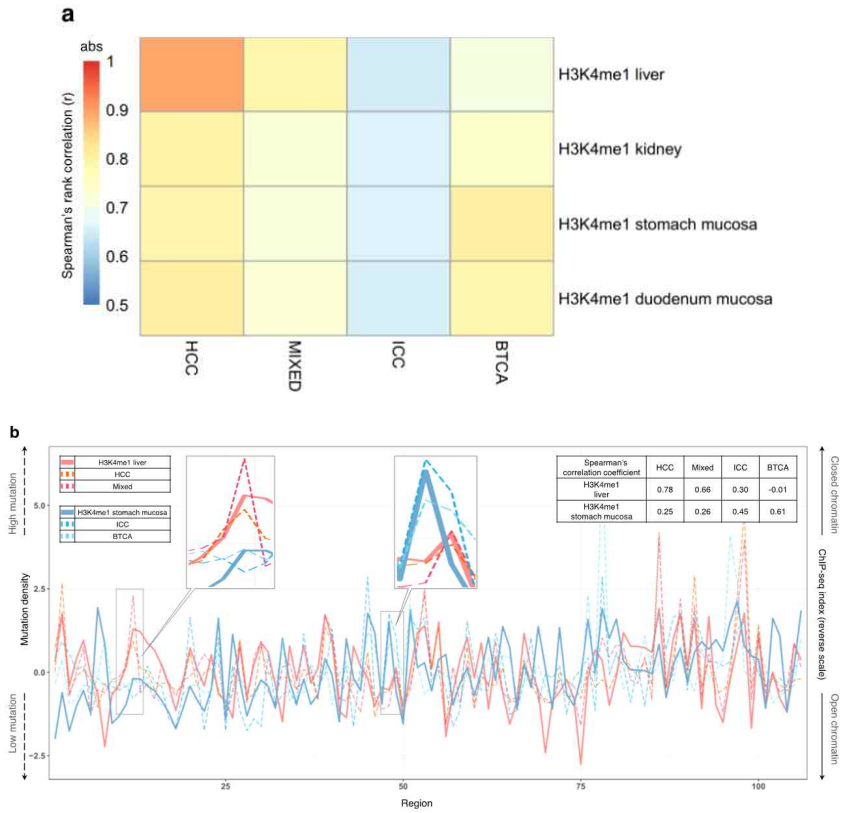
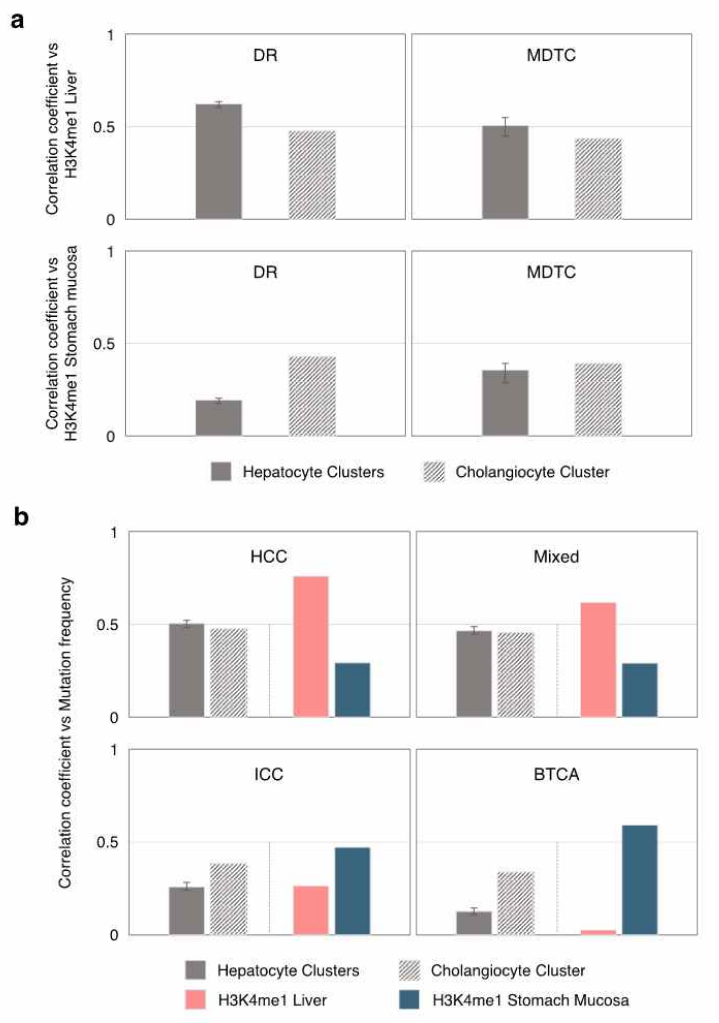


Figure S9. Correlations between cancer genome mutation density and the H3K4me1 chromatin features in different tissue types. (a) Correlations between somatic mutation frequencies and chromatin landscape in all 2128 regions represented by heat map. Different color depths correspond to the absolute values 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).


a
Correlation coefficient vs H3K4me1 Liver
DR
MDTC
Correlation coefficient vs H3K4me1 Stomach mucosa
Hepatocyte Clusters
Cholangiocyte Cluster
b
Correlation coefficient vs Mutation frequency
HCC
Mixed
ICC
BTCA
Hepatocyte Clusters
Cholangiocyte Cluster
H3K4me1 Liver
H3K4me1 Stomach Mucosa

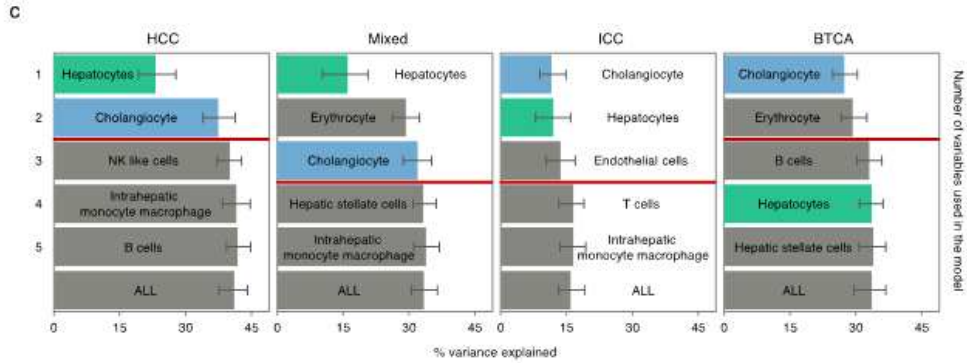

c
HCC
Mixed
ICC
BTCA
Hepatocytes
Cholangiocyte
NK like cells
Intrahepatic monocyte macrophage
B cells
ALL
Erythrocyte
Hepatic stellate cells
Endothelial cells
T cells
% variance explained
Number of variables used in the model

Figure S10. Analysisamong regional somatic mutation frequencies, H3K4me1 chromatin features and scRNA-seq gene expression factor levels. (a, b) Correlation coefficient values from 106 1Mbp sub-selected regions displaying the largest differences in the regression model between H3K4me1 liver and stomach mucosa were calculated using Spearman's rank method. In the case of hepatocyte clusters, averaged correlation values obtained by assessing the correlations between DR or MDTC of each hepatocyte cluster and either chromatin features or regional somatic mutation frequencies were used. Minimum and maximum values are represented as the error bars. (a) Bar graphs representing correlation coefficients between H3K4me1 chromatin features from either liver (upper part) or stomach mucosa tissue (bottom part) and DR or MDTC factor levels of hepatocyte clusters or cholangiocyte cluster. (b) Bar graphs demonstrating correlation coefficients between the sub-selected regional mutation frequency of each PLC subtype and DR values from cell clusters (left part of each inset figure) or H3K4me1 chromatin marks from two tissues (right part of each inset figure). Each inset figure corresponds to each PLC subtype. All correlation values were converted to absolute values for visual purposes. (c) Random forest regression-based scRNA-seq gene expresssion factor feature selection employing aggregated mutation frequency from each PLCs subtype. The average DR values of clusters were used for calculating scRNA-seq feature of each cell type. The rank of each scRNA-seq feature is estimated by importance values. The bar length shows the variance explained scores, and error bar indicates minimum and maximum scores derived from 1,000 repeated simulations. Red lines show the cutoff scores determined by the prediction accuracy of total features-1 s.e.m. hepatocyte scRNA-seq feature is green-colored and cholangiocyte scRNA-seq feature is blue-colored.

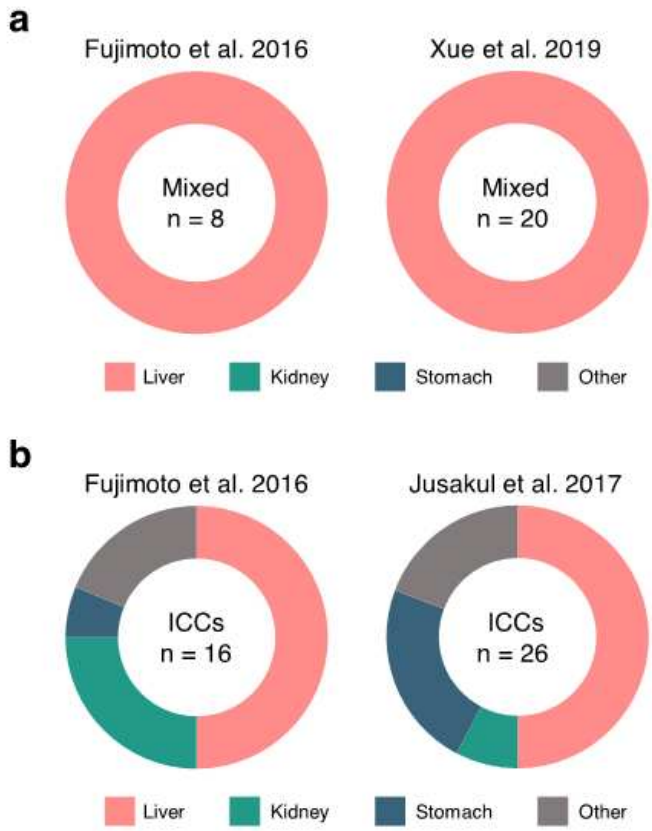


Figure S11. Cell-of-origin prediction distributions for distinct Mixed and ICCs cohorts using chromatin features. Pie graphs represent the percentage of samples with COO assignments as liver tissue chromatin features (pink), kidney tissue chromatin features (green), stomach tissue chromatin features (navy) or the rest (gray). (a) Comparison between Mixed subtype cohorts and (b) between ICC subtype cohorts.

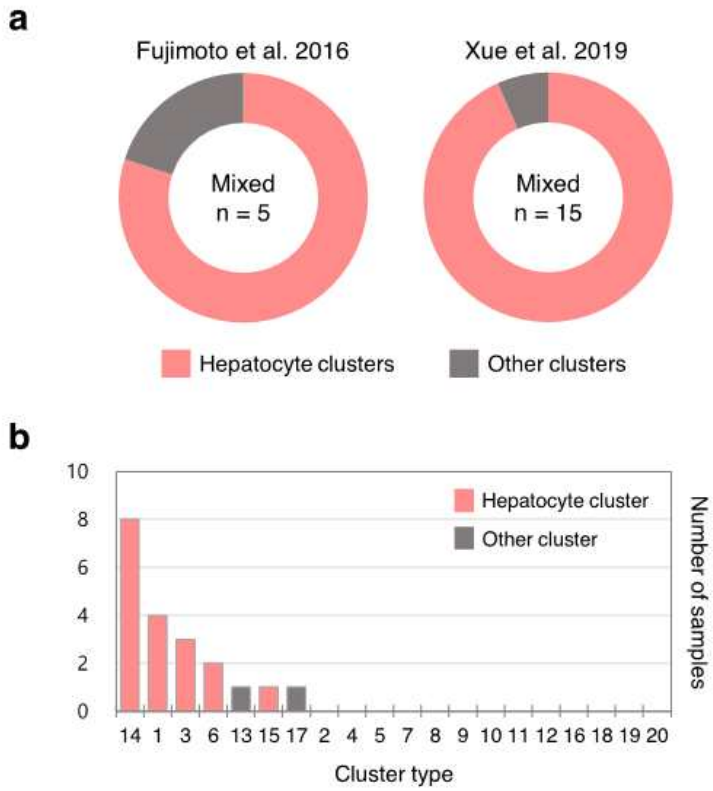


Figure S12. Cell-of-origin prediction using scRNA-seq data. scRNA-seq features derived from a total of 20 single cell clusters constituting human liver tissue was employed to elucidate the relationship with the regional mutation frequency of Mixed type at individual sample level. (a) COO prediction for two distinct Mixed subtype cohorts. Pie graphs indicate the percentage of COO assignments as hepatocyte clusters (pink) or other clusters (gray). (b) Bar graph displaying the number of Mixed subtype samples from the two cohorts assigned per each cluster. Hepatocyte clusters are pink-colored and other clusters are gray-colored.

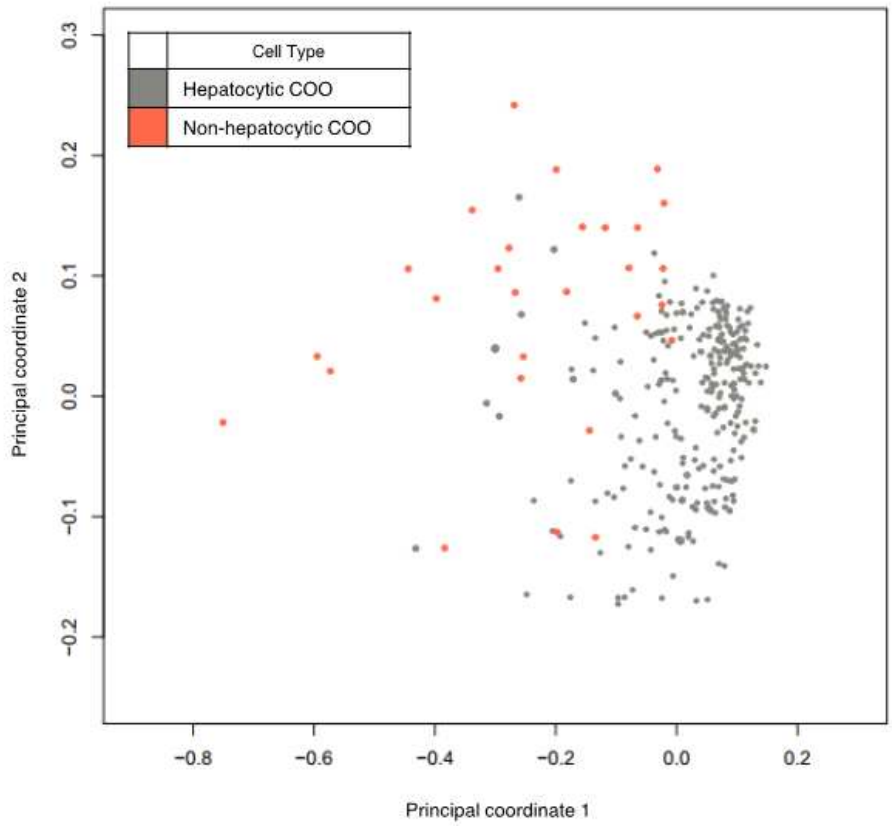


Figure S13. PCOA of individual cancer samples. Hepatocytic COO samples are gray-colored and non-hepatocytic COO samples are orange-colored.

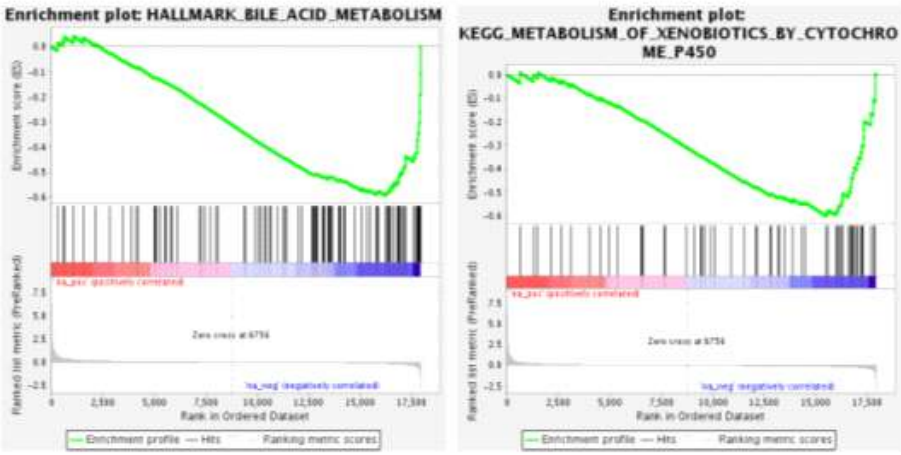


Figure S14. Gene sets that were down-regulated in non-hepatocytic COO HCCs.

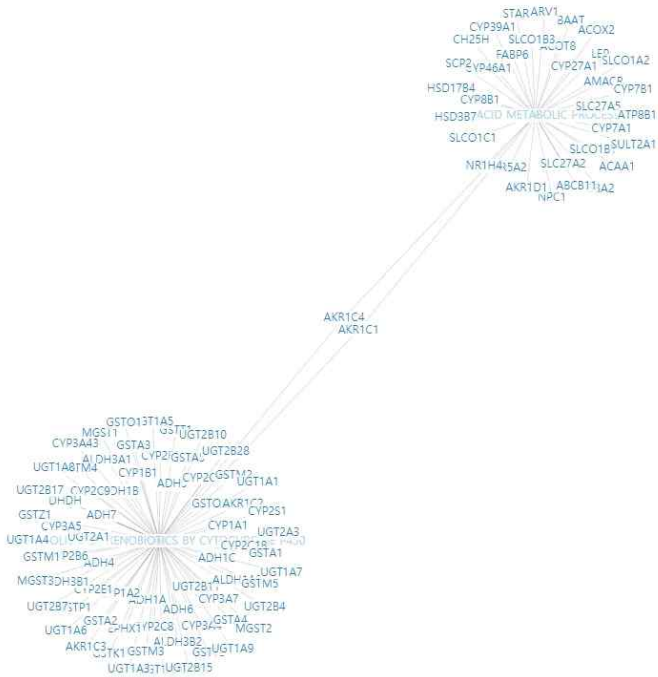


Figure S15. Connection network of 'bile acid synthesis' and 'xenobiotic degradation' pathway through Bio-Entity Explorer.

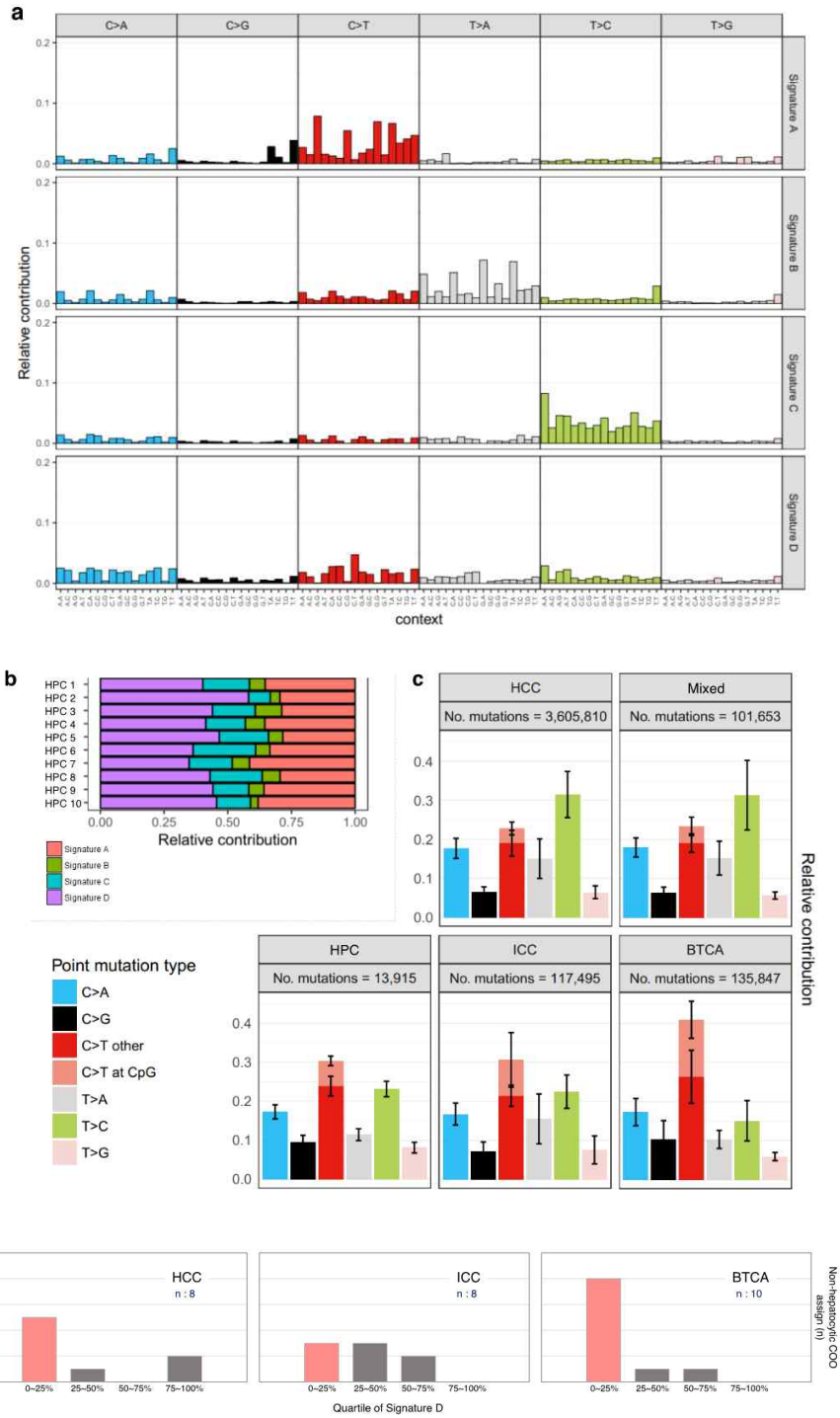

a
C>A
C>G
C>T
T>A
T>C
T>G
Signature A
Signature B
Signature C
Signature D
Relative contribution
context
b
HPC 1
HPC 2
HPC 3
HPC 4
HPC 5
HPC 6
HPC 7
HPC 8
HPC 9
HPC 10
Relative contribution
Signature A
Signature B
Signature C
Signature D
c
HCC
No. mutations = 3,605,810
Mixed
No. mutations = 101,653
HPC
No. mutations = 13,915
ICC
No. mutations = 117,495
BTCA
No. mutations = 135,847
Relative contribution
Point mutation type
C>A
C>G
C>T other
C>T at CpG
T>A
T>C
T>G
d
HCC
n : 8
ICC
n : 8
BTCA
n : 10
0~25%
25~50%
50~75%
75~100%
Quartile of Signature D
Non-hepatocytic COO assign (n)

Figure S16. 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.

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

Reference

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Borresen-Dale, A.L., et al. (2013). Signatures of mutational processes in human cancer. Nature 500, 415-421.

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., and Cuppen, E. (2018). MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Med 10, 33.

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., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechno l31, 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.

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.

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

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

Hayakawa, Y., Sethi, N., Sepulveda, A.R., Bass, A.J., and Wang, T.C. (2016). Oesophageal adenocarcinoma and gastric cancer: should we mind the gap? Nat Rev Cancer 16, 305-318.

Hodgkinson, A., Chen, Y., and Eyre-Walker, A. (2012). The large-scale distribution of somatic mutations in cancer genomes. Hum Mutat 33, 136-143.

Innes, B.T., and Bader, G.D. (2018). scClustViz - Single-cell RNAseq cluster assessment and visualization. F1000Res7.

Jung, J.-u., Lim, J.-M., Joe, H., and Kim, H.-G. (2020). BEE : a web service for biomedical entity exploration. bioRxiv.

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. Cancer Discov 7, 1116-1135.

Kan, Z., Jaiswal, B.S., Stinson, J., Janakiraman, V., Bhatt, D., Stern, H.M., Yue, P., Haverty, P.M., Bourgon, R., Zheng, J., et al. (2010). Diverse somatic mutation patterns and pathway alterations in human cancers. Nature 466, 869-873.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature 502, 333-339.

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.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al.(2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499, 214-218.

Liu, L., De, S., and Michor, F. (2013). DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. Nat Commun 4, 1502.

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.

Martincorena, I., and Campbell, P.J. (2015). Somatic mutation in cancer and normal cells. Science 349, 1483-1489.

Matsumoto, T., Takai, A., Eso, Y., Kinoshita, K., Manabe, T., Seno, H., Chiba, T., and Marusawa, H. (2017). Proliferating EpCAM-Positive Ductal Cells in the Inflamed Liver Give Rise to Hepatocellular Carcinoma. Cancer Res 77, 6131-6143.

Michalopoulos, G.K., Barua, L., and 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., 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 Hepato l66, 952-961.

Monga, S.P. (2019). Updates on hepatic homeostasis and the many tiers of hepatobiliary repair. Nat Rev Gastroenterol Hepato l16, 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., 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.
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.
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.
Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and 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.
Ross-Innes, C.S., Becq, J., Warren, A., Cheetham, R.K., Northen, H.,

O'Donovan, M., Malhotra, S., di Pietro, M., Ivakhno, S., He, M., et al. (2015). Whole-genome sequencing provides new insights into the clonal architecture of Barrett's esophagus and esophageal adenocarcinoma. Nat Genet 47, 1038-1046.

Schaefer, M.H., and Serrano, L. (2016). Cell type-specific properties and environment shape tissue specificity of cancer genes. Sci Rep 6, 20707.

Schuster-Bockler, B., and Lehner, B. (2012). Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature 488, 504-507.

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

Sekiya, S., and Suzuki, A. (2014). Hepatocytes, rather than cholangiocytes, can be the major source of primitive ductules in the chronically injured mouse liver. Am J Patho l184, 1468-1478.

Stamatoyannopoulos, J.A., Adzhubei, I., Thurman, R.E., Kryukov, G.V., Mirkin, S.M., and Sunyaev, S.R. (2009). Human mutation rate associated with DNA replication timing. Nat Genet 41, 393-395.

Supek, F., and Lehner, B. (2015). Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature 521, 81-84.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature 489, 75-82.

Tummala, K.S., Brandt, M., Teijeiro, A., Grana, O., Schwabe, R.F., Perna, C., and 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., and Nusse, R. (2015). Self-renewing diploid Axin2(+) cells fuel homeostatic renewal of the liver. Nature 524, 180-185.

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 Hepatol68, 959-969.

Woo, Y.H., and Li, W.H. (2012). DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. Nat Commun 3, 1004.

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. Cancer 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., and 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. Cancer Cell 23, 784-795.
Zhang, L., Zhou, Y., Cheng, C., Cui, H., Cheng, L., Kong, P., Wang, J., Li, Y., Chen, W., Song, B., et al. (2015). Genomic analyses reveal mutational signatures and frequently altered genes in esophageal squamous cell carcinoma. Am J Hum Genet 96, 597-611.

국문초록

식도암과 간암의 기원세포 다양성에 대한 연구

하경식
서울대학교 의료정보학 협동과정

다중 조직 기반의 하위 유형을 가지고 있는 원발성 식도암 및 간암은 전 세계의 질병 부담 및 사망률을 지속적으로 증가시키고 있다. 이러한 원발성 암에 대한 기원 세포의 규명은 각 암 유형별로 관련 치료제 및 예방 약품을 보다 넓은 폭으로 선택할 수 있는 기회를 제공할 수 있다. 지금까지 각 암종과 그들이 포함하고 있는 하위 유형에 대한 기원 세포의 이질성을 다루는 여러 연구들이 있었지만, 인체 유래 세포를 사용하여 각 암에 대한 기원 세포를 추적하는 연구는 제대로 수행되지 않았다. 우리는 종양 조직과 정상 조직의 후성유전학적 표지에 대한 전체 게놈 시퀀싱 데이터를 분석하여 각 원발성 암의 하위 유형에 대한 기원 세포 예측을 수행하였다. 특히, 간암의 경우에 인체유래의 single cell RNA-seq 데이터를 분석하여 기원세포 예측의 정확도를 높이는데 집중하였다. 우리의 분석 결과 식도선암종과 식도편평세포암종이 동일한 조직에서 암이 발생했음에도 불구하고 두 암종의 기원세포는 서로 다를 수 있음을 보여주었고, 특히 선암종의 경우에는 대부분의 샘플들이 위 세포에서 유래된 것으로 나타났다. 우리의 분석은 또한 간암에서 드물게 발생하는 하위 유형이며 또한 간세포와 담관세포가 혼합되어 있는 조직학적 특징을 가지고 있는 혼합 간세포-담관암의 유래가 간세포 기원인 것을 밝혀내었다. 그리고 간세포 암종에서 드물게 비 간세포 기원으로 예측된 샘플들의 경우에는 담관 세포 특이적 마커인 *EPCAM*의 높은 발현량이 나타나기도 하였다. 추가적으로 간 전구 세포의 전체 게놈 시퀀싱 데이터를 분석하여 이러한 전구 세포가 간암의 직접적인 기원 세포가 아닐 수 있

다는 가능성을 확인하기도 하였다. 종합적으로 이러한 결과들은 원발성 암 기원 세포의 다양성에 대한 새로운 통찰력을 제공할 것으로 기대된다.

키워드: 체세포돌연변이, 후성유전학, 기원세포, 기계학습, 식도암, 간암
학 번: 2011-23816