HHS Public Access

Author manuscript

Cancer Discov. Author manuscript; available in PMC 2018 April 01.

Published in final edited form as:

Cancer Discov. 2017 October ; 7(10): 1116–1135. doi:10.1158/2159-8290.CD-17-0368.

Whole-Genome and Epigenomic Landscapes of Etiologically Distinct Subtypes of Cholangiocarcinoma

A full list of authors and affiliations appears at the end of the article.

Abstract

Cholangiocarcinoma (CCA) is a hepatobiliary malignancy exhibiting high incidence in countries with endemic liver-fluke infection. We analysed 489 CCAs from 10 countries, combining whole-genome (71 cases), targeted/exome, copy-number, gene expression, and DNA methylation information. Integrative clustering defined four CCA clusters – Fluke-Positive CCAs (Clusters 1/2) are enriched in *ERBB2* amplifications and *TP53* mutations, conversely Fluke-Negative CCAs (Clusters 3/4) exhibit high copy-number alterations and *PD-1*/*PD-L2* expression, or epigenetic mutations (*IDH1/2*, *BAP1*) and *FGFR*/*PRKA*-related gene rearrangements. Whole-genome analysis highlighted *FGFR2* 3′UTR deletion as a mechanism of *FGFR2* upregulation. Integration of non-coding promoter mutations with protein-DNA binding profiles demonstrates pervasive modulation of H3K27me3-associated sites in CCA. Clusters 1 and 4 exhibit distinct DNA hypermethylation patterns targeting either CpG islands or shores – mutation signature and subclonality analysis suggests that these reflect different mutational pathways. Our results exemplify how genetics, epigenetics and environmental carcinogens can interplay across different geographies to generate distinct molecular subtypes of cancer.

Keywords

Cholangiocarcinoma; Liver Fluke; Whole genome Sequencing; Methylation; Biliary tract cancer

INTRODUCTION

Cholangiocarcinoma (CCA) is the second most common hepatobiliary malignancy, accounting for 10–20% of primary liver cancers (1). The highest rates of CCA are in South East Asia (Northeast Thailand, Cambodia, and Laos), where >8,000 cases are diagnosed annually due to infection by liver flukes such as *Opisthorchis viverrini* and *Clonorchis sinensis* (2). CCA is considered relatively rare in Western countries (<6 per 100,000 population) where risk factors such as primary sclerosing cholangitis, hepatolithiasis and

Corresponding Authors: Patrick Tan, Duke-NUS Medical School, 8 College Road, Singapore 169857; Phone: 65-6516-1783; Fax: 65-6221-2402; gmstanp@duke-nus.edu.sg; Bin Tean Teh, Duke-NUS Medical School, 8 College Road, Singapore 169857; Phone: 65-6601-1324 teh.bin.tean@singhealth.com.sg; Chawalit Pairojkul, Faculty of Medicine, Khon Kaen University, Khon Kaen, Thailand 40002; Phone: 66-43363691; Fax: 66-43348388; chawalit-pjk2011@hotmail.com; Tatsuhiro Shibata, National Cancer Center Research Institute, Tokyo, Japan 1040045; Phone: 81-335422511; Fax: 81-335475137; tashibat@ncc.go.jp; Steven G. Rozen, Duke-NUS Medical School, 8 College Road, Singapore 169857; Phone: 65-9857-3213; Fax: 65-6534-8632; steve.rozen@duke-nus.edu.sg; and Raluca Gordân, Department of Biostatistics and Bioinformatics, Department of Computer Science, Duke University, Durham, North Carolina, USA 27708; Phone: 1-919-6849881; Fax: 1-919-6680695; raluca.gordan@duke.edu.

*Contributed equally to this work

Conflict of interest: The authors declare no potential conflicts of interest.

choledochal cysts predominate (3). However, the incidence of intrahepatic CCA in the USA appears to be increasing (1).

Current 5-year survival rates for CCA after surgery and chemotherapy remain poor [<20%, (4)], and clinical trials evaluating targeted therapies in unselected CCA populations have shown minimal benefits (5). Existing CCA classification systems are primarily based on either anatomical location (intrahepatic, perihilar, and distal) or pathological features (cirrhosis, viral hepatitis, and primary sclerosing cholangitis), which do not provide insights into mechanisms of CCA tumorigenesis, nor potential targets for therapy. While previous exome-sequencing studies by our group and others have revealed a complex CCA mutational landscape (6–8), no study to date has compared fluke-positive and fluke-negative CCAs at the whole-genome level, nor fully explored the extent and contribution of structural variants and non-coding regulatory mutations to CCA pathogenesis. Little is also known about epigenetic differences between fluke-positive and fluke-negative CCAs.

Here, on behalf of the International Cancer Genome Consortium, we report an integrated genomic, epigenomic and transcriptomic analysis of CCA involving nearly 500 CCAs from ten countries. Comprehensive integrative clustering revealed four CCA clusters likely driven by distinct etiologies, with separate genetic, epigenetic, and clinical features. Our analysis uncovered new driver genes (*RASA1, STK11, MAP2K4, SF3B1*) and structural variants (*FGFR2* 3′ UTR deletion). Within the non-coding genome, we observed a significant enrichment of promoter mutations in genes regulated by epigenetic modulation, uncovered through a novel analysis framework incorporating experimentally-derived protein-DNA binding affinities and pathway information. Strikingly, we found that two of the CCA clusters displayed distinct patterns of DNA hypermethylation enriched at different genomic regions (CpG islands vs shores), demonstrating for the first time the existence of distinct DNA methylation subgroups of CCA. We propose that tumors from these two subtypes may have arisen through distinct mechanisms of carcinogenesis, driven by either extrinsic carcinogenic agents or intrinsic genetic insults.

RESULTS

CCA Whole-Genome Sequencing and Integrative Clustering

Whole-genome sequencing (WGS) was performed on 71 CCA tumors and non-malignant matched tissues, including both fluke-associated (Fluke-Pos, 22 *O. viverrini* and 1 *C. cinensis* samples) and non-fluke associated cases (Fluke-Neg, 48 samples). Of these, 27 samples have been previously analysed at the exome level (6–8) (Supplementary Tables 1A–B). Sequencing was performed to an average depth of 64.2× (median 65.1×; Supplementary Table 1C). We called somatic mutations (single-nucleotide variants (sSNVs) and short insertion-deletions (indels)) using both Genome Analysis Toolkit and MuTect (Supplementary Methods). Orthogonal validation resequencing on 97 randomly selected sSNVs and 85 indels using either Ion Torrent or Sanger technologies determined accuracy rates to be 99% for sSNVs and 87% for indels. In total, we detected 1,309,932 somatic mutations across the 71 tumors, with 4,541 nonsilent sSNVs and 1,251 nonsilent indels in protein-coding genes. On average, each CCA had 82 nonsilent somatic mutations per tumor (median 47), consisting of 64 nonsilent somatic sSNVs (median 41) and 18 indels (median

6). These mutation counts are comparable to those previously observed in genomes from pancreatic cancer (9) (74 sSNVs and 5 indels per tumor), liver cancer with biliary phenotype (10) (79 sSNVs and 24 indels per tumor) and hepatocellular carcinoma (11) (70 sSNVs and 6 indels per tumor). Three CCAs exhibited exceptionally high mutation levels (average 5.91 sSNVs/Mb and 24.17 indels/Mb, vs. 1.39 and 3.72 for other CCAs) – these tumors exhibited mutational signatures of microsatellite instability (MSI) and two of these cases exhibited *PolE* mutations. Excluding these three hypermutated cases, Fluke-Pos CCAs exhibited significantly more somatic mutations compared to Fluke-Neg CCAs (median 4,700 vs. 3,143 per tumor, $p < 0.05$, Wilcoxon rank-sum test).

Previous studies by our group have suggested that genomic alterations in CCA may differ according to causative etiology (6–8). However, definitive exploration of these differences has been missing, due to limitations in sample sizes and genes analysed, reliance on a single genomic platform (exome), and lack of whole-genome information. To address these limitations, we assembled a cohort of 489 CCAs (Supplementary Tables 1A–B) including 133 Fluke-Pos (132 *O. viverrini* and 1 *C. cinensis* samples) and 356 Fluke-Neg cases. Samples were analysed using four different genomic platforms based on sample availability. Besides WGS (71 cases), these included exome sequencing of 200 cases (previously published (8)), high-depth targeted sequencing of 188 cases, SNP array copy-number profiling of 175 cases, array-based DNA methylation profiling of 138 cases, and array-based expression profiling of 118 cases. To confirm the applicability of merging different datatypes, we used statistical models to confirm that our mutation calls were not biased by differences in sequencing platforms, and directly confirmed mutation concordance by analysing those samples sequenced on overlapping platforms (Supplementary Methods). Using iClusterPlus, we performed integrative clustering combining data from somatic mutations, somatic copy-number alterations (sCNAs), mRNA expression and DNA methylation on 94 CCAs where all four data types were available. Randomized subsampling clustering confirmed the robustness of these integrative clusters (Supplementary Fig. 1A). To support the reliability of our conclusions, reanalysis using an expanded set of integrative clustered samples (121 samples), including samples with one or more missing platforms while retaining a cluster prediction accuracy of 90%, yielded similar results and associations (Supplementary Fig. 1B, Supplementary Methods). We also found that while the most discriminatory clustering was achieved by using all available genomic information, clustering by individual modalities recapitulated some of the integrative clusters, for sCNA, expression, and methylation data (Supplementary Fig. 1C, see later section on DNA methylation). Importantly, integrative clustering performed on samples stratified by anatomical location reproduced the original clusters within each anatomical site (≥ 88% concordance; Supplementary Fig. 1D), demonstrating that the molecular clusters are not simply recapitulating anatomical variation.

Integrative clustering revealed 4 distinct clusters characterized by different clinical features and genomic alterations (Fig. 1A). Cluster 1 comprised mostly Fluke-Pos tumors, with hypermethylation of promoter CpG islands (aberrantly methylated above normal level, Supplementary Methods), enrichment of *ARID1A* and *BRCA1/2* mutations ($p < 0.01$ and $p < 0.05$ respectively, Fisher's exact test), high levels of nonsynonymous mutations ($p < 0.001$, Wilcoxon rank-sum test, Supplementary Fig. 2A), and high levels of mutations in gene

promoters with histone 3 lysine 27 trimethylation (H3K27me3) predicted to alter transcription factor binding (see later section on somatic promoter mutations). Cluster 2 was characterized by a mix of Fluke-Pos and Fluke-Neg tumors, with upregulated *CTNNB1*, *WNT5B* and *AKT1* expression ($p < 0.05$, Wilcoxon rank-sum test, Supplementary Fig. 2B) and downregulation of genes involving *EIF* translation initiation factors (Supplementary Table 2A). Most notably, Clusters 1 and 2 were also significantly enriched in *TP53* mutations and *ERBB2* amplifications ($p < 0.001$ and $p < 0.01$ respectively, Fisher's exact test) and elevated *ERBB2* gene expression ($p < 0.05$, Wilcoxon rank-sum test, Fig. 1A and Supplementary Fig. 2C).

In contrast to Clusters 1 and 2, Clusters 3 and 4 comprised mostly Fluke-Neg tumors. Cluster 3 displayed the highest level of sCNAs, including enrichment of amplifications at chromosome arms 2p and 2q ($q < 0.05$, Fisher's exact test; Supplementary Table 2B). Analysis of immune populations by ESTIMATE revealed that both Clusters 2 and 3 displayed immune cell infiltration (Supplementary Fig. 2D), but only Cluster 3 exhibited specific upregulation of immune checkpoint genes (*PD-1*, *PD-L2* and *BTLA*) (Fig. 1B and Supplementary Fig. 2E) and pathways related to antigen cross-presentation, CD28 co-stimulation, and T cell signal transduction (Supplementary Table 2A). Cluster 4 was characterized by *BAP1*, *IDH1/2* mutations, *FGFR* alterations (all $p < 0.01$, Fisher's exact test), and upregulated *FGFR* family and *PI3K* pathway signatures (Fig. 1A and Supplementary Table 2A). Similar to Cluster 1, Cluster 4 tumors also exhibited DNA hypermethylation – however, rather than hypermethylation at CpG islands, Cluster 4 hypermethylation was at CpG promoter shores (see later section on DNA methylation).

We sought to relate the CCA clusters to anatomical and clinical features. Clusters 1 and 2 were enriched in extrahepatic (consisting of perihilar and distal) tumors while Clusters 3 and 4 were composed almost entirely of intrahepatic tumors ($p < 0.001$, Fisher's exact test). This was observed in both fluke-negative and fluke-positive tumors, and persisted after adjusting for fluke status ($p < 0.001$, multivariate regression; also confirmed in expanded clusters). Other CCA risk factors, such as Hepatitis B Virus (HBV), Hepatitis C Virus (HCV), and primary sclerosing cholangitis (PSC), were present in our cohort at frequencies of 10.4%, 2.9%, and 1.0% respectively. Both HBV and PSC were associated with intrahepatic CCA ($p < 0.05$, Fisher's exact test) (12).

We further investigated the prevalence of driver genes according to anatomical location (among 459 samples with sequencing and anatomical location information). *BAP1* and *KRAS* were more frequently mutated in intrahepatic cases ($q < 0.1$, Fisher's exact test). This was observed in both fluke-negative and fluke-positive tumors, and persisted even after adjusting for fluke status ($q < 0.1$, multivariate regression). No additional genes were identified as differentially mutated when the extrahepatic CCAs were analysed as perihilar and distal CCAs.

Clinically, patients in Clusters 3 and 4 had significantly better overall survival relative to the other 2 clusters ($p < 0.001$, log-rank test; Fig. 1C). As fluke infection was also associated with poorer survival ($p < 0.001$, log-rank test; Supplementary Fig. 2F), we performed multivariate analysis and confirmed that this cluster-associated survival difference persisted

even after accounting for fluke association, anatomical location, and clinical staging ($p < 0.05$, Cox proportional hazards model; Supplementary Table 2C). To validate this finding, we assembled a separate validation cohort comprising newly-classified samples from the expanded integrative clustering and a recently-published set of CCA samples (13) (Supplementary Methods). A survival analysis on this independent validation cohort reaffirmed the same survival trends, on both univariate ($p < 0.05$, log-rank test; Supplementary Fig. 2F) and multivariate analysis ($p < 0.05$, Cox proportional hazards model; Supplementary Table 2C). This result demonstrates that molecular clusters can provide additional prognostic information in a manner independent of fluke status and anatomical location. Fig. 1D summarizes the salient features of each CCA cluster.

New CCA Driver Genes and Structural Rearrangements

Driver gene mutation analysis across 459 CCAs (130 Fluke-Pos and 329 Fluke-Neg cases) revealed 32 significantly mutated genes (SMGs, $q < 0.1$ by both MutSigCV and IntOGen; Supplementary Tables 3A–D and Fig. 2A). Our analysis revealed four potentially new CCA driver genes not highlighted in previous CCA publications (6–8,13–17) – *RASA1, STK11, MAP2K4*, and *SF3B1*. Of these, *RASA1, STK11*, and *MAP2K4* are related to RAS/MAPK signalling (Fig. 2B–C and Supplementary Table 3B). *RASA1*, encoding a p120 Ras GTPase-activating protein, was predicted to be inactivated in 4.1% of cases (10 frame shift, 4 nonsense; Fig. 2D). These inactivating mutations, along with observed focal *RASA1* copy-number losses, were associated with decreased *RASA1* expression (Fig. 2E). In CCA cell lines, shRNA-mediated knockdown of *RASA1* resulted in significantly enhanced migration and invasion, supporting a tumor-suppressor role for *RASA1* in CCA (Fig. 2F). *STK11*, a serine/threonine protein kinase, was mutated in 5% of cases, with most *STK11* mutations also predicted to be inactivating (7 nonsense, 9 frame shift; Fig. 2G). *SF3B1*, an RNA splicing factor, was mutated in 4.6% of cases, at mutation hotspots (23% at codon 625 and 14% at codon 700) previously observed in uveal melanoma and breast cancer (18,19) (Fig. 2H). This latter finding may implicate a role for RNA splicing dysregulation in CCA tumorigenesis. *MAP2K4* is a member of the mitogen-activated protein kinase, and has been shown to activate p38 mitogen-activated protein kinase and c-Jun N-terminal kinase. We observed *MAP2K4* focal homozygous deletions in two Fluke-Pos cases (Supplementary Fig. 3A–C), and *MAP2K4* mutations in another 10 cases (2.2%). Half of these were predicted to be inactivating (frameshift, nonsense, splice site mutations), consistent with a tumor suppressor role for *MAP2K4* in CCA.

ERBB2 was amplified in 3.9 – 8.5% of CCAs (Supplementary Table 3E). *ERBB2* amplifications were more frequent in Fluke-Pos cases (10.4% in Fluke-Pos vs 2.7% in Fluke-Neg CCA, $p < 0.01$, Fisher's exact test) with elevated *ERBB2* gene expression in Fluke-Pos compared to Fluke-Neg tumors in these cases (Supplementary Fig. 2C, Supplementary Table 3E for validation samples). On average, *ERBB2* amplified samples exhibited 14 *ERBB2* copies (copy numbers determined by ASCAT (SNParray) or Quandico (sequencing data)), and GSEA analysis confirmed upregulation of *ERBB2*-related gene sets among *ERBB2*-amplified samples ($q < 0.2$; Supplementary Table 3F). We independently validated the presence of ERBB2 amplification in selected cases by flouresence *in-situ* hybridization (FISH; Supplementary Fig. 3D). Other pathways upregulated in *ERBB2*-

amplified samples included biological oxidation, metabolism cytochrome P450, peroxisome proliferator-activated receptors signaling and checkpoint signaling ($q < 0.2$). In addition to *ERBB2* amplifications, we also detected activating *ERBB2* mutations (S310F/Y, G292R, T862A, D769H, L869R, V842I, G660D) in 9 cases (2%). Notably, previous studies in cell lines have shown that high *ERBB2*-expressing CCAs may be more sensitive to ERBB2-inhibitor treatment compared to low *ERBB2*-expressing cases, suggesting that tumors with high *ERBB2* expression may be candidates for anti-ERBB2/HER2 therapy (20). Amplification of other selected oncogenes included *MYC* (n = 12), *MDM2* (n = 9), *EGFR* (n = 11), and *CCND1* (n = 7), while deletions included *CDKN2A* (n = 17), *UTY* (n = 17), and *KDM5D* (n = 16) (Supplementary Table 3G).

Availability of WGS data also allowed us to investigate the role of SVs in CCA. Using CREST, we identified ~93 somatic SVs per tumor (median 69, range 0 to 395), with a 91% (61/67) true positive rate by PCR. Most of the SVs were intra-chromosomal (65%), and associated with cancer-related genes (*ARID1A, CDKN2A/B*), retrotransposon-associated genes (*TTC28*), and fragile sites (1q21.3) (Supplementary Table 4). SV burden varied significantly across the 4 CCA clusters ($p < 0.05$, Kruskal-Wallis test), with Fluke-Neg tumors in Cluster 4 associated with low burden ($p < 0.05$, 1-sided Wilcoxon rank-sum test). *TP53*, *FBXW7* and *SMAD4* were significantly associated with increased SV burden ($q < 0.1$, Wilcoxon rank-sum test; Supplementary Fig. 4A).

FGFR2 fusion genes, previously reported in CCA (21), are thought to deregulate FGFR2 signalling through the hijacking of 3′ fusion partners with dimerization motifs (21,22). However, whether rearrangements affecting other *FGFR* members (besides *FGFR2*) exist in CCA, or if other categories of *FGFR2* rearrangements (besides in-frame gene fusions) can contribute to CCA development, remain unclear. Analysis of the WGS data, followed by subsequent validation at the gene transcript level, revealed 5 in-frame gene fusions with intact tyrosine kinase domains – four involving *FGFR2* (*FGFR2-STK26, FGFR2-TBC1D1, FGFR2-WAC*, and *FGFR2-BICC1)* (Fig. 3A) and one involving *FGFR3 (FGFR3-TACC3)* (Fig. 3B). To our knowledge, this is the first report of *FGFR3* fusions in CCA. *FGFR3-TACC3* fusions have been previously reported in bladder cancer, glioblastoma, and lung cancer (23,24) and were shown to be oncogenic.

Besides *FGFR* in-frame fusions, we also identified recurrent truncating events translocating *FGFR2*, without its 3′ UTR, to intergenic regions (Fig. 3C). *FGFR2*-truncated CCAs exhibited high expression levels compared to *FGFR2* transcripts with intact 3′UTRs ($p < 0.01$, Wilcoxon rank-sum test). *In vitro*, luciferase reporter experiments confirmed diminished expression in constructs containing *FGFR2* 3′UTRs compared to control reporters (Fig. 3D). *FGFR2* 3′UTR loss may thus represent a new and additional mechanism for enhancing *FGFR2* expression in CCA, similar to mechanisms reported for *PD-L1* (25).

FGFR2 rearrangements were observed exclusively in Cluster 4 ($p < 0.001$, Fisher's exact test) (Fig. 1A). In CCAs lacking *FGFR* rearrangements, we further identified other genomic alterations involving *FGFR2* including indels (n=3), SNVs (n=10), and copy-gains (n=1) – several of these alterations have been previously shown to be activating (26,27). Collectively, CCAs with altered *FGFR2* genes (point mutations, indels, copy-gain and rearrangements)

were significantly enriched in Cluster 4 ($p < 0.01$, Fisher's exact test) and also expressed significantly higher *FGFR2* levels than *FGFR2*-wildtype tumors (Fig. 1A, Fig. 2B and Fig. 3E).

In addition to *FGFR* fusions, WGS analysis also identified rearrangements affecting the catalytic subunit B of cAMP-dependent protein kinase A (*PRKACB*), including *ATP1B1-PRKACB* and *LINC00261-PRKACB* (Fig. 3F). Both *PRKACB* rearrangements retained the *PRKACB* pseudokinase domain, and thus may increase PKA activity and activate downstream MAPK signalling (8).

Long interspersed nuclear element-1 (LINE-1 or L1) repeats are autonomous retrotransposons collectively occupying ~17% of the human genome. Recent studies have shown that certain L1 elements are active in cancer, displaying somatic retrotransposition and potentially contributing to genomic instability and tumorigenesis (28–30). In the CCA WGS samples, we observed frequent somatic L1 retrotranspositions, particularly originating from an L1 element in intron 1 of the *TTC28* gene (52 events in 20/71 tumors, 28.2%, Supplementary Fig. 4B). PCR testing validated 98% (46/47) of these retrotransposition events (Supplementary Table 4). CCAs with L1 retrotransposition were associated with Fluke-Pos tumors ($p < 0.01$, Fisher's exact test) and increased SV burden ($p < 0.05$, Wilcoxon rank-sum test), suggesting a relationship between genomic instability and L1 endonuclease activity. Further analysis revealed that these intragenic insertions were not overtly associated with cancer-related genes, such as tumor suppressors.

CCA Somatic Promoter Mutations

Somatic mutations in noncoding regulatory regions have been proposed to play crucial roles in carcinogenesis (31). However, systematic approaches for identifying such mutations are lacking and their effects in CCA remain poorly understood. To date, only *TERT* promoter mutations have been observed in CCA (8), and in our WGS cohort, only two CCAs (2.8%) harboured *TERT*-promoter mutations (chr5:1295228). Besides *TERT*, no other recurrent non-coding promoter-region mutations were observed in the WGS cohort.

Even when integrating mutations over regulatory regions or gene promoters, the low recurrence rate of most noncoding mutations may lead to a lack of statistical power to identify potential drivers in the promoters of individual genes. We hypothesized that the effects of noncoding promoter mutations might instead be detectable at the gene-set level. To test this hypothesis, we developed a novel method, FIREFLY (FInding Regulatory mutations in gEne sets with FunctionaL dYsregulation, Fig. 4A), which identifies gene sets dysregulated by somatic promoter mutations that alter transcription factor (TF) binding. Compared to approaches used in previous cancer genome studies (32,33), FIREFLY differs in three important respects. First, it uses experimentally determined high-throughput TF-DNA binding data for 486 TFs representing a broad range of TF families (34,35), as opposed to position weight matrices (PWMs) (36), to predict mutation-associated changes in TF binding affinity. Second, FIREFLY condenses the large numbers of highly non-recurrent noncoding mutations into biologically-meaningful gene sets, shortlisting those sets with an overrepresentation of mutations, as assessed by multiple statistical tests. Third, it

orthogonally validates the transcriptional consequences of the binding-change predictions using expression data from primary tumors.

To identify sets of genes that had dysregulated transcription in the aggregate due to promoter mutations, we applied FIREFLY to 70 WGS samples (1 hypermutated sample was excluded), representing 6,639 somatically mutated gene promoters. Based on binding-change predictions, FIREFLY identified 138 sets of genes that were enriched for binding-change mutations in promoters ($q < 0.1$, Fisher's exact test with Benjamini-Hochberg false discovery rate). FIREFLY's second statistical test then compared the number of binding-change mutations in these sets to an expected null distribution determined from synthetically mutated data, created using the tumor-specific mutation rates for each type of mutation in its trinucleotide context. Nineteen sets passed the second test ($q < 0.1$). Finally, FIREFLY orthogonally assessed the transcriptional impact of the binding-change predictions, by testing whether tumors with increasing numbers of binding-change mutations for a given gene set also exhibit greater transcriptional dysregulation in that set ($q < 0.1$, by Gene Set Analysis (37)). Four of the 19 sets were validated by this test (Fig. 4B; see Fig. 4C and Supplementary Fig. 5A for example non-significant gene sets). This was significantly greater than expected for a null distribution based on randomly selected gene sets of similar sizes ($p < 0.01$; Supplementary Fig. 5B). We note that FIREFLY results do not imply that every somatic promoter mutation in the 4 gene sets affects gene expression. Instead, the results indicate that the identified sets have excesses of promoter mutations that likely affect gene expression. We also note that FIREFLY does not use predicted gain or loss of binding to infer directionality of change in expression levels. This is because, in general, TFs can act as either activators or repressors depending on the regulatory context. Consequently, one cannot in general predict whether a gain or loss of a binding site will result in up or down-regulation.

To validate a fundamental assumption in the FIREFLY pipeline, i.e. that mutations predicted to alter TF binding also affect transcription, we selected three mutations and tested them in luciferase reporter assays. We confirmed altered regulatory activity for two of the three mutations, providing direct evidence that mutations predicted to change TF binding can in fact alter gene expression (Supplementary Fig. 5C).

FIREFLY identified 4 gene sets that were likely dysregulated by altered TF-DNA binding. Interestingly, two of these (MIKKELSEN_MCV6_HCP_WITH_H3K27ME3 and MIKKELSEN_MEF_ICP_WITH_H3K27ME3) are subsets of PRC2 target genes that have promoters with histone modification H3K27me3 in certain contexts (Fig. 4B). Mutations in all 4 enriched gene sets occurred across all four of the CCA clusters. It was noteworthy, however, that Cluster 1 was enriched for mutations in 3 of the 4 gene sets, including two associated with H3K27me3 (Supplementary Fig. 5D). This, together with the observed hypermethylation of Polycomb Repressive Complex 2 (PRC2) target genes in Cluster 1 (see later section on DNA methylation), provides additional evidence of the importance of alterations in PRC2 regulation in Cluster 1.

Distinct CCA Epigenomic Subtypes

As shown in Fig. 1, iCluster analysis revealed two distinct hypermethylated CCA subgroups (Clusters 1 and 4). To validate these differences across a larger CCA series, we performed unsupervised DNA-methylation clustering on an expanded panel of 138 CCAs. Clustering based solely on DNA methylation recapitulated two hypermethylated clusters highly concordant with Clusters 1 and 4 (96.3% and 86.1% concordance respectively; Fig. 5A), and a third group of low-methylation tumors representing a mix of Clusters 2 and 3. Cluster 1, enriched in Fluke-Pos CCAs, was dominated by hypermethylation in promoter CpG islands, while Cluster 4, enriched in Fluke-Neg CCAs, was dominated by hypermethylation in promoter CpG island shores (Fig. 5B). Different sets of gene promoters were targeted for hypermethylation in Clusters 1 and 4. However, Gene Set Enrichment Analysis (GSEA) revealed that both affected common pathways, including PRC2 targets. We observed significant inverse correlations between transcript levels and promoter methylation in both Cluster 1 and 4 ($q < 0.05$; Supplementary Fig. 6A), consistent with these epigenetic alterations exerting transcriptional impact.

DNA hypermethylation in the two hypermethylated clusters may be driven by distinct epigenetic mechanisms. In Cluster 1, we observed downregulation of the DNA demethylation enzyme *TET1* and upregulation of the histone methyltransferase *EZH2* (Supplementary Fig. 6B), suggesting a possible role for these genes in establishing the hypermethylation phenotype (38). In contrast, Cluster 4 CCAs were significantly enriched in *IDH1/2* mutations, which are known to be associated with CCA hypermethylation (7,13,39) (31.6% in Cluster 4 versus 1.0% in other clusters, $q < 0.001$, multivariate regression; Fig. 5A). Among Cluster 4 CCAs lacking *IDH1/2* mutations (68.4%), *BAP1* mutations were enriched ($q < 0.001$ and 0.05 respectively for inactivating point mutations and regional deletions; Fig. 5A). *BAP1* mutated cases were also associated with increased CpG hypermethylation relative to *BAP1* wildtype cases (Supplementary Fig. 6C). Notably, *BAP1* mutations have been associated with DNA hypermethylation in CCA and renal cell carcinoma (13,40).

To explore mutation patterns between these two clusters, we identified ten established mutation signatures in the WGS cohort. These included COSMIC Signatures 1, 5, 8, 16, and 17, and signatures associated with activated APOBECs (Signatures 2 and 13), mismatch-repair deficiency (MMR, Signatures 6 and 20), and aristolochic acid exposure (Signature 22, Supplementary Fig. 6D). Signature 5 burdens were correlated with patient age (Spearman correlation 0.25, $p < 0.05$) as previously reported for other cancer types (41). Fluke-Pos CCAs were enriched for activated APOBEC mutation burden ($p < 0.001$, multivariate regression). Signatures of MMR and Signatures 8, 16 and 17 have not been previously reported in CCA.

We observed elevated levels of Signature 1 (CpG>TpG mutations) in Cluster 1, even after adjusting for patient age ($p < 0.001$, multivariate regression; Fig. 5C). Importantly, this elevation is Signature 1-specific, as it was not observed for Signature 5. We note that CpG dinucleotides are known mutation hotspots, due to spontaneous deamination of 5-methylcytosine (5mC) to thymine (CpG>TpG mutation) (42). To investigate if hypermethylated CpGs in Cluster 1 might provide susceptible genomic substrates for

deamination and subsequent Signature 1 mutations, we integrated the locations of the CpG>TpG mutations with regions of hypermethylation. In Cluster 1, CpG>TpG mutations were indeed located preferentially near hypermethylated regions ($p < 0.001$, Fisher's exact test; Fig. 5D and Supplementary Fig. 6E), whereas in Cluster 4 these mutations showed no such regional preferences. These results support a significantly increased level of DNA hypermethylation-related deamination events in Cluster 1 compared to Cluster 4.

We further investigated if the differences in genome-wide mutation patterns between clusters are accompanied by distinct clonal structures harbouring these mutations. Distribution analysis of variant allele frequencies (VAFs) for point mutations (in copy-neutral regions and adjusted for tumor purity) revealed a wide spread of VAFs in Cluster 1 compared to Cluster 4 (Fig. 5E), indicating the presence of heterogeneous subclones in Cluster 1 tumors, but not in Cluster 4 tumors. Together, these distinct patterns of hypermethylation-related deamination and tumor heterogeneity suggest disparate somatic-evolution processes during tumorigenesis in these clusters (see Discussion).

DISCUSSION

Surgery is the only proven treatment modality for CCA (43), and all formal evaluations of targeted therapies to date, performed in unselected CCA cohorts, have proved unsuccessful (44). In this study, we analysed a cohort of nearly 500 CCAs from distinct geographical regions, including 94 CCAs covered by four genomic platforms. Integrative clustering of mutation, copy number, gene expression, and epigenetic data revealed four subtypes of CCA, each exhibiting distinct molecular and clinicopathologic features. Four lines of evidence highlight that clustering based on molecular profiles provides additional information beyond anatomical site. Firstly, anatomical site does not drive molecular subtypes, as evidenced by the reproducibility of the molecular subtypes within each anatomical site separately (Supplementary Fig. 1D). Secondly, tumors in different anatomical sites may exhibit similarities at the molecular level, while tumors located in the same anatomical site can display profound differences in their molecular profiles. This is exemplified by Clusters 1 and 2 comprising mixtures of intrahepatic and extrahepatic tumors, while intrahepatic tumors are split among all 4 clusters. Thirdly, from a disease prognosis standpoint, CCAs in different anatomical sites do not differ in their survival trends, whereas the molecular clusters show significant differences in survival, in both the original and an independent validation cohort (Fig. 1C and Supplementary Fig. 2F). Finally, we note that current medical oncology guidelines do not discriminate CCA treatments based on their anatomical site (1) as classifications based on anatomy do not provide information regarding potential therapeutic opportunities. In contrast, the molecular profiles highlight several potential cluster-specific therapies (see next paragraph). Taken collectively, our results indicate that molecular CCA subtypes based on integrative molecular clustering are likely to offer enhanced information regarding CCA biology and clinical behaviour, beyond that provided by anatomical location alone.

Examination of signature genomic alterations in each subtype highlights potential therapeutic opportunities, although we emphasize that such findings require further clinical validation. For example, it is possible that CCAs in Clusters 1 and 2 with *ERBB2*

amplification may be appropriate for therapies targeting ERBB2/HER2 signaling (45). The elevated expression of immune related genes and pathways in Cluster 3 CCAs suggests a therapeutic opportunity for immunotherapy, however this conclusion should be treated with caution due to the small sample size of this cluster. The specific mechanisms driving the elevated expression of immune-related genes in Cluster 3 remain unclear, however increased levels of immunogenicity and upregulation of MHC protein and antigen processing complexes have been independently observed in aneuploid tumors (46,47), consistent with the elevated level of chromosomal aberrations in Cluster 3. Cluster 4 CCAs, which are associated with *IDH1/2* mutations and *FGFR2* and *PRKA*-related gene rearrangements, might also be tested with recently described IDH inhibitors (ClinicalTrials.gov identifier: NCT02073994) or FGFR-targeting agents. Notably, our study suggests that loss of *FGFR2*-3′ UTRs may represent an alternative mechanism of *FGFR2* activation, beyond *FGFR2* in frame gene fusions and activating mutations. This, along with our additional observation of *FGFR3* rearrangements, expands the proportion of potential FGFR-targetable cases in CCA (48).

Previous cancer studies have discovered recurrent regulatory mutations in individual promoters such as *TERT* (33,49). In complement to these studies, we developed in this work an alternative analysis framework (FIREFLY), which uses protein binding microarray (PBM)-based analysis to examine the effects of non-coding promoter mutations on gene-expression pathways in cancer. FIREFLY's use of PBM data with a *k*-mer-based approach provides advantages for estimating the effects of mutations on TF binding affinitites compared to traditional PWM models. For example, PWMs implicitly assume that each base pair contributes independently to binding affinity, however this is not always true. Moreover, PWMs do not capture the multiple modes of binding associated with many TFs, and PWM scores are not easily comparable between TFs. FIREFLY addresses these shortcomings by using experimentally determined PBM data, which provides actual binding affinities of every possible 8-mer. Using FIREFLY, we identified several pathways with strong statistical evidence for recurrent systematic dysregulation in CCA. Interestingly, two of these pathways reflected epigenetically-modulated cellular differentiation processes, which have been shown to be dysregulated in cancer in general by other means such as DNA hypermethylation or histone modification.

Of the four clusters, two clusters (Cluster 1 and 4) are most clearly distinguished by their highly distinctive patterns of genome-wide DNA hypermethylation, targeting either promoter CpG islands or promoter CpG shores. To our knowledge, such epigenetically distinct tumor subtypes have not been previously reported in the literature, particularly for tumors from the same tissue-type. Further analysis demonstrated that Cluster 1 CCAs are Fluke-Pos with elevated mutation rates, Mutation Signature 1 enrichment, and increased point-mutation subclonality, while Cluster 4 CCAs are Fluke-Neg and by comparison relatively clonal. We propose that these differences are consistent with a model where early in tumorigenesis, Cluster 1 CCAs are likely driven by external carcinogenic agents and early epigenetic deregulation ("epimutations"), while Cluster 4 CCAs are likely driven by pioneer genetic events such as *IDH1/2* or *BAP1* mutations, with epigenetic aberrations arising as a downstream consequence (Fig. 6). In this model, fluke infection, by inducing chronic inflammation, metabolic disruption of host bile homeostasis (50), or secretion of fluke

growth factors and excretory vesicles for modulating host-pathogen interactions (51), induces genome-wide epigenetic deregulation. Cytosine residues experiencing aberrant methylation are then at higher risk of spontaneous deamination and mutation, consistent with the enrichment of Signature 1 mutations in this subtype. CpG island hypermethylation in this subtype may also silence tumor suppressor genes, further enhancing cancer development. Moreover, because the processes of carcinogen-induced methylation, deamination, and mutation are inherently stochastic from cell to cell, such events would inevitably lead to increased levels of intra-tumor heterogeneity. In contrast, in Cluster 4 CCAs which are Fluke-Neg, somatic mutations in critical chromatin modifier genes (e.g. *IDH1/2* and *BAP1*) may occur as a primary event preceding epigenetic deregulation, driving both rapid clonal outgrowth and directly inducing DNA hypermethylation. Specifically, *IDH1/2* mutations have been shown to increase 2-hydroxyglutarate oncometabolite production, leading to DNA hypermethylation.

We acknowledge that other models may also explain the striking molecular differences between Cluster 1 and 4. These include differential vulnerabilities in distinct cells-of-origin, as the biliary system is known to contain multipotent stem/progenitor cells. Liver fluke infection primarily affects large intra and extra- hepatic bile ducts, giving rise to intra and/or extra-hepatic CCA. Conversely, parenchymal liver diseases exclusively affect canals of Hering and bile ductules, and are primarily associated with intrahepatic CCA (52).

A recent study from the USA TCGA consortium reported a multi-omic analysis of a smaller and more-homogenous CCA series (38 samples, exclusively fluke-negative, mostly intrahepatic and North American) (13). Comparison of this study to our own data revealed an obvious difference – specifically our inclusion of Fluke-Pos samples allowed the discovery of another major CCA subtype (Cluster 1) with distinct clinical and molecular features. Besides Cluster 1, the other study's "IDH" (*IDH* mutants) and "METH3" (*BAP1* mutants and *FGFR* rearrangements) groups are likely matches to our Cluster 4 (characterized by *IDH* and *BAP1* mutants, and *FGFR* rearrangements); while their "ECC" group (extrahepatic) matched Cluster 2 (containing fluke-negative extrahepatic tumors). On the other hand, the TCGA "METH2" group (*CCND1* amplifications) and our Cluster 3 were not obviously matched. Taken collectively, these results suggest that most of the TCGA study's clusters are largely concordant with our own, and neither classification strictly precludes the other.

We acknowledge that limitations of sample resources (DNA, RNA and paraffin-embedded tissues) were a major constraint in this study. We were unable to generate data using all platforms on all samples, which reduced the sample size in the integrative clustering analysis. To overcome this constraint, we sequenced an extended and separate sample cohort, to validate findings emerging from the integrative clustering. Centre-specific differences in pre-sample processing steps, including collection site, biopsy site, and sample processing protocols, may also result in sequencing biases. We attempted to control for these variations by reviewing the histology of all cases using standardized AJCC 7th criteria to confirm and harmonize the histology and anatomical subtype of samples. We also attempted to reduce biopsy bias and normal contamination originating from the biopsy sites by estimating tumor cell content through histopathological review or SNP arrays. Lastly, our

study merged data from different DNA sequencing platforms (WGS, WES, and targeted sequencing), thus limiting our analysis across the entire cohort to genomic regions common to these platforms. However, we overcome biases from different data processing centers and merging different data types by using one analysis pipeline, ensuring uniform analysis on each data type, and confirming good concordance between samples sequenced on multiple platforms (Supplementary Methods).

In summary, integrative analysis of a large CCA cohort has revealed a novel molecular taxonomy, with discovery of new potential CCA driver genes and gene rearrangements. This taxonomy may be clinically relevant, however more functional data is required to test the validity of the genomic data. Analysis of non-coding promoter mutations, made possible by whole-genome analysis, revealed that they play a significant role in CCA targeting genes involved in PRC2 biology, and we identified two highly distinct CCA subtypes demonstrating distinct DNA hypermethylation patterns. We conclude by noting that these last two findings may carry conceptual relevance beyond biliary tract cancers. Specifically, while elevated DNA methylation levels have been observed in numerous cancers, these cases to date been largely consigned to a general CIMP (CpG Island Methylator Phenotype) class – our data suggests that a more detailed examination of these cases may reveal potential for epigenetic heterogeneity. Finally, as exemplified by the *TERT* promoter, previous analyses of non-coding promoter mutations have largely focused on identifying individual regions of recurrent genomic aberration. Our FIREFLY analysis suggests that such cases are likely rare, and that similar to protein coding genes, mutations in non-coding regulatory regions may target genes in specific pathways rather than individually. Extending this concept to promoter mutation catalogues of other cancer types will certainly test the general applicability of this proposition.

METHODS

Sample collection

Primary tumor and matched normal samples (non-neoplastic liver or whole blood) were obtained from the SingHealth Tissue Repository (Singapore), Fundeni Clinical Institute (Romania), Khon Kaen University (Thailand), ARC-Net Biobank (Italy), Centre de Ressources Biologiques Paris-Sud (France), Department of Pathology, Yonsei University College of Medicine (South Korea), Hospital do Cancer de Barretos (Brazil), Linkou Chang Gung Memorial Hospital (Taiwan) and Department of Pancreatobiliary Surgery, The First Affiliated Hospital, Sun Yat-Sen University (China) with signed informed consent. WGS data from 10 Japanese cases were also included (National Cancer Center, Japan). The study was approved by the SingHealth Centralised Institutional Review Board (2006/449/B), Ethics Committee of the Clinical Institute of Digestive Diseases and Liver Transplantation, Fundeni (215/18.01.2010), Khon Kaen University (HE471214), Centre de Ressources Biologiques Paris-Sud, National Cancer Center, Japan (G20-03), Severance Hospital, Yonsei University Health System (4-2014-0829), Hospital do Cancer de Barretos (716/2013), Linkou Chang Gung Memorial Hospital (100-2030B), The First Affiliated Hospital of Sun Yat-Sen University (2014/C_006) and ARC-Net Biobank at Verona University Hospital (n. prog. 1959). Clinico-pathological information for subjects including age, sex, histology,

tumor subtype, stage and overall survival were reviewed retrospectively. Cases were staged according to the American Joint Committee on Cancer (AJCC) Staging System 7th Edition. All patients had not received prior treatment. In total, 489 tumors with associated clinicopathologic data were obtained (133 Fluke-Pos: 132 *O. viverrini*, 1 *C. sinensis*; 39 HBV/HCV-positive; 5 PSC-positive). These were assayed on at least one profiling platform, which included: (1) WGS (71 CCAs); (2) targeted sequencing surveying 404 genes (188 CCAs); (3) published exome sequencing ((8); 200 CCAs); (4) HumanOmniExpress BeadChip arrays (SNP arrays; 175 CCAs); (5) DNA methylation 450k BeadChip arrays (138 CCAs) and (6) HumanHT-12 Expression BeadChip arrays (Gene expression arrays; 118 CCAs) (Supplementary Table 1 and Supplementary Table 5). A detailed list of clinical data is included in Supplementary Table 1A.

Cell lines

H69 non-malignant immortalized cholangiocyte cells were obtained in 2011, from D. Jefferson (New England Medical Center, Tufts University) and cultured as previously described (53). HEK293T cells were obtained from ATCC (CRL-3216) in 2015 and cultured with Dulbecco's Modified Eagle's Medium, 10% Fetal Bovine Serum, 2mM L-glutamine and 1% Penicillin/Streptomycin. EGI-1 was purchased from DSMZ (ACC 385) and maintained in DMEM supplemented with 10% FBS (Sigma-Aldrich). HUCCT1 (JCRB0425) cells were purchased from the Health Sciences Research Resources Back (HSRRB) in 2009 and maintained in RPMI1640 medium with 10% FBS. M213 (JCRB1557) cells were obtained from the Liver Fluke and Cholangiocarcinoma Research Center in 2015 and cultured with Ham's F12 media (Gibco). Cells were cultured at 37 °C in a 5% CO2 humidified chamber. All cell lines were authenticated by short-tandem repeat (STR) profiling and found to be negative for mycoplasma as assessed by the MycoSensor qPCR Assay Kit (Agilent Technologies).

Sample preparation and sequencing

Genomic DNA was extracted using the QIamp DNA mini kit (Qiagen). DNA yield and quality were determined using Picogreen (Invitrogen) and further visually inspected by agarose gel electrophoresis. RNA was extracted using an RNeasy mini kit (Qiagen), quantified by measuring Abs260 with a UV spectrophotometer, and quality assessed with the Agilent 2100 Bioanalyzer (Agilent Technologies). Sequencing libraries were prepared from DNA extracted from tumor and normal samples using the SureSelect XT2 Target Enrichment System for the Illumina Multiplexed Sequencing platform (Illumina) according to the manufacturer's instructions. Whole-genome sequencing was performed using Illumina Hiseq X10, Illumina Hiseq2500 and Illumina Hiseq2000 instruments. To survey the frequency and distribution of somatic mutations in the validation cohort (188 CCAs), targeted sequencing of 404 genes was performed after capture with a custom SureSelect capture reagent designed using the SureDesign tool (Agilent Technologies). Target-enriched libraries were sequenced on the Illumina HiSeq 4000 sequencing platform. Coverage of coding regions based on the amplicon design was 99.6% (Supplementary Table 5).

Reporter assays

For 3′ UTR Reporter Assays, control reporter plasmids (LUC) were generated by cloning the SV40 promoter into the pGL4.10 promoterless luciferase reporter vector (Promega). The LUC-FGFR2_3′UTR test plasmid was engineered by inserting the 1,666 bp 3′UTR region of *FGFR2* (starting from stop codon) into the immediate 3′ end of luciferase gene of the LUC plasmid. HEK293T or H69 cells were cotransfected with a Renilla-containing plasmid with either LUC or LUC-FGFR2_3′UTR. Transfected cells were incubated for 24 hours.

For promoter mutation reporter assays, Luciferase constructs were generated by ligating fragments of promoter regions (2 kb upstream and 500 bp downstream of the TSS) prepared from genomic PCR using gene-specific primers. Mutant constructs were generated using the QuikChange II XL site-directed mutagenesis kit (Agilent Technologies) according to the manufacturer's instructions. H69 or EGI-1 cells were transfected and incubated for 48 hours, after which Dual-Luciferase Reporter Assays (Promega) were performed. Relative luciferase activity was calculated as the firefly luciferase activity normalized to Renilla (Promega) luciferase activity. Assays were conducted in triplicates and each experiment was repeated 3 independent times.

RASA1 shRNA silencing and Functional Assays

Construction of hairpin-pLKO.1 vectors (carrying a puromycin antibiotic resistance gene) containing shRNAs targeting *RASA1* coding sequences were performed as follows: shRNA#1 (CCGGGCTGCAAGAACACTGATATTACTCGAGTAATATCAGTGTTCTTGCAGCTTTTTG, TRCN0000356449, Sigma) and shRNA#2 (CCGGCCTGGCGATTATTCACTTTATCTCGAGATAAAGTGAATAATCGCCAGGTTTTT, TRCN0000005998, Sigma). Lentivirus particles were produced in HEK-293T cells transfected with psPAX2, PMD2G (Addgene) and pLKO.1-shRNA-containing plasmids. M213 and HUCCT1 cells were infected with the lentivirus, and stable cells were established by puromycin selection (Sigma). Migration assays were performed in Transwells (Corning Inc., 8.0-μm pore size). For migration, 2.5×10^4 cells of *RASA1* stably-silenced M213 and HUCCT1 cells in serum-free medium were added to 24-well insert plates. Media containing 10% FBS were added to the lower wells and incubated for 5 h. For cell invasion assays, the filters were pre-coated with Matrigel. 2.5×10^4 cells of *RASA1* stably-silenced M213 and HUCCT1 cells in serum-free medium were added into 24-well insert plates. Media containing 10% FBS was added to the lower well of the chambers and incubated for 5 h. After incubation, the cells on the upper surface of the filter were completely removed and migrated cells were trypsinized and counted.

Somatic mutation detection

We aligned sequence data to the human reference genome (hs37d5) using BWA-MEM v0.7.9a (54). We removed PCR duplicates using SAMTools (55) and performed indel realignments using Genome Analysis Toolkit v1.0 (GATK) (56). We used realigned data as input to both GATK Unified Genotyper and MuTect to call sSNVs (somatic single nucleotide variants). We used GATK IndelGenotyperV2 to identify indels. We applied filters

and manual inspection to retain only high confidence sSNVs and indels. Further details are provided in the Supplementary Methods.

Analysis of somatic promoter mutations with FIREFLY

FIREFLY (FInding Regulatory mutations in gEne sets with FunctionaL dYsregulation) is a method for identifying gene sets dysregulated by somatic promoter mutations through modulation of TF binding. We extracted somatic promoter mutations by selecting non-coding sSNVs within +/− 2kb of TSSs of GENCODE genes. We identified those mutations predicted to change TF-binding based on PBM data (details below). We then evaluated gene sets for enrichment in promoter binding-change mutations. For each gene set, we calculated the test statistic M = number of genes in the gene set with binding-change mutations, summed across all tumors. To identify gene sets enriched in binding-change mutations, we performed two statistical tests in sequence, where gene sets found significant in each test ($q < 0.1$) were then tested in the next: (1) Fisher's exact test; (2) a synthetic mutations test. For gene sets passing these tests, we then performed gene expression tests using the "GSA" v1.03 R package (37) to assess their transcriptional dysregulation. Further details are provided in the Supplementary Methods.

We generated synthetic mutations in promoter regions of interest based on the genome-wide frequency of mutations at each trinucleotide observed for each tumor/normal pair. Synthetic mutations were generated separately for each tumor, based on the frequencies and mutation types observed in the actual tumor. Further details are provided in the Supplementary Methods.

Identifying TF binding-change mutations

To identify binding-change mutations, we used TF-DNA binding specificity PBM data from the cis-BP database (34). PBM experiments measure TF binding to all possible 8-mer sequences with a DNA binding enrichment score (E-score) (35). Typically, E-scores > 0.35 correspond to specific TF-DNA binding (57). To call binding sites for a particular TF, we required that such sites contain at least two consecutive 8-mers with E-scores > 0.4. We also used PBM data to call 'non-binding sites' defined as genomic regions containing only 8-mers with E-score < 0.3. For each somatic mutation and each TF with available PBM data, we analyzed the 15-bp genomic region centered at the mutation. If the region contains a TF binding site in the normal sample but not in the corresponding tumor sample, the mutation was called a 'loss-of-binding' mutation for that TF. If the region contains a TF binding site in the tumor sample but not in the normal sample, then the mutation was called 'gain-of-binding' for that TF. We describe a mutation as 'binding-change' if it is either loss-of-binding or gain-of-binding for any of the interrogated TFs. Further details are provided in the Supplementary Methods.

Driver gene analysis

We integrated (i) 71 WGS CCAs, (ii) 188 targeted-sequenced CCAs (Supplementary Table 5), and (iii) 200 exome-sequenced CCA (8), and performed gene significance analyses using MutSigCV (58) and IntOGen (59). For input, we used the list of all coding sSNVs and indels found in the 404 targeted genes across 459 samples (including both silent and

nonsilent mutations). Both tools were run with default parameters and we retained genes found significant by both tools with q values <0.1. Further details are provided in the Supplementary Methods.

Detection and annotation of structural variations

BWA-MEM alignments from each tumor-normal pair were analyzed by CREST (Clipping REveals STructure) (60) and PTRfinder (61). For most tumours we required ≥ 3 uniquely mapped split-read alignments at each SV breakpoint; for the shallower Japanese WGS data we required only ≥ 5 such alignments over both SV breakpoints combined. We considered a tumor SV to be somatic if no SV in the normal sample occurred within one-half of a read length from the tumor SV. Further details are provided in the Supplementary Methods.

Identification of L1-retrotransposition insertions

We searched for sources of somatic L1 insertions by looking for highly recurrent SVs: ≥ 10 SVs in a 1Mb region. We then selected the subset of these region that contained a mobile L1 element in a database of retrotransposon insertion polymorphisms (dbRIP) (62). Only SVs with ≥ 2 reads with poly-A tails at the putative L1 insertion site were retained for further analysis.

Validation of structural variations

For genomic DNA, 100 ng of whole-genome amplified DNA of the tumor and normal matched cases were used as PCR templates. For cDNA, total cDNAs of tumor and normal matched control were synthesized using SuperScript III System according to manufacturer's instructions (Invitrogen) and 40 ng of cDNA were used as PCR template. PCR was performed using fusion-specific primers with Platinum Taq DNA Polymerase system (Invitrogen). PCR products were cleaned up by the Exo/Sap enzyme system (Invitrogen) and bidirectionally sequenced using the BigDye Terminator v.3.1 kit (Applied Biosystems) and an ABI PRISM 3730 Genetic Analyzer (Applied Biosystems). Sequencing traces were aligned to reference sequences using Lasergene 10.1 (DNASTAR) and analysed by visual inspection.

Somatic L1 insertions were validated by PCRs using primers flanking predicted sites of insertion. PCRs were performed using AccuPrime™ Pfx DNA Polymerase (Invitrogen) with 200 ng of WGA DNA.

Copy-number analysis

Raw SNP array data was processed using Illumina Genome Studio. We used ASCAT v2.0 (63) to estimate allele-specific copy-number profiles. We determined regions of copy-number alteration based on their relative copy-number using the "copynumber" R package (64). For tumor-normal pairs without SNP array data, we estimated copy-number profiles based on sequencing data using Control-FREEC (65), Quandico (66), and Sequenza (67). We used GISTIC v2.0.22 (68) to determine regions of significant focal copy-number alterations, using ASCAT/Sequenza's inferred copy-number segments, and associated copy-number values were defined as $\log_2$ of the segment's relative copy-number. Full details are provided in the Supplementary Methods.

Gene expression analysis

Gene expression microarray data was pre-processed using the “lumi” R package (69). Batch effects were removed using ComBat (70). We used GSEA v2.2.2 (71) with a classic weighting scheme to determine pathways upregulated or downregulated in each integrative CCA cluster relative to the others, employing canonical pathways in the MSigDB C2 catalogue of annotated gene sets. Full details are provided in the Supplementary Methods.

Immune cell infiltration analysis

ESTIMATE (72) was used to determine the presence of infiltrating immune cells, using the ImmuneSignature geneset. A total of 126 genes were used to determine the immune score for each tumor.

DNA methylation analysis

DNA methylation profiles were obtained for 138 tumors and 4 normal samples. Data was preprocessed using the “minfi” (73) and “wateRmelon” (74) R packages. We selected 4,520 probes with the highest standard deviations in β-values across the tumors, and mean $\beta < 0.5$ in the normal samples, for clustering using the “RPMM” R package (75). In the hypermethylated methylation clusters (1 and 4), we considered a CpG site to be hypermethylated if the following conditions held: (1) $\beta < 0.5$ in normal samples; (2) M-values were significantly different in the (i) hypermethylated cluster versus (ii) the combined normal samples and the low-methylation tumors—those not in methylation cluster 1 or 4 ($q < 0.05$, two-sided t-test); and (3) its mean β in the hypermethylated cluster minus the mean β across the normal samples and low-methylation tumors was > 0.2.

To explore associations between mutation signatures and hypermethylated CpGs, we considered only mutations located within 50 bp of CpG probes that had mean $\beta < 0.5$ in normal samples. In each tumor, the nearest CpG probe to a mutation was considered to be hypermethylated if it was: (1) hypermethylated in that tumor’s methylation cluster; (2) its individual β was > 0.5; and (3) its individual β minus the mean β across the normal samples and low-methylation tumors was > 0.2. Other analyses were similar to community-standard analyses. Further details are provided in the Supplementary Methods.

Integrative clustering

The “iClusterPlus” R package (76) was used to perform integrative unsupervised clustering of 94 CCAs based on 4 genomic data types: (i) somatic point mutations in 404 targeted genes (gene by sample matrix of binary values), (ii) sCNAs defined as copy-number segments identified by ASCAT v2.0, (iii) the most variable expression probes (coefficient of variation > 0.1), and (iv) the most variable methylation probes (top 1% standard deviation in β-value). We ran iClusterPlus.tune with different numbers of possible clusters (n=2 to 7), choosing the number of clusters at which the percentage of explained variation leveled off (n=4), and the clustering with the lowest Bayesian information criterion (BIC). Further details are provided in the Supplementary Methods.

Survival analysis

The “survival” R package was used to perform survival analysis using Kaplan-Meier statistics, with p-values computed by log-rank tests. Multivariate survival analysis was performed using the Cox proportional hazards method. To validate our survival analysis results, we also analysed a separate validation cohort of 58 samples, by combining two sources: (i) 25 samples (with survival data) newly classified into CCA clusters under the expanded integrative clustering; and (ii) 33 recently-published CCA samples with survival data from Farshidfar et al. (2017) (13). Further details are provided in the Supplementary Methods.

Mutation signature analysis

Non-negative matrix factorization (NMF) was applied to the trinucleotide-context mutation spectra of CCAs to extract mutation signatures. Six stable and reproducible extracted mutational signatures were compared to the 30 signatures from the Catalogue of Somatic Mutations in Cancer (COSMIC) (http://cancer.sanger.ac.uk/cosmic/signatures) based on cosine similarities. We used supervised NMF to evaluate the contributions of the COSMIC signatures to mutations in CCA. Additional signatures were considered via visual inspection. We ignored signatures that contributed < 5% of the total mutations in a particular tumor, and removal of these signatures did not substantially increase reconstruction errors. The MSI status in the prevalence set was determined by the indel counts in simple repeat sequences. Further details are provided in the Supplementary Methods.

Additional information

Accession codes—The whole genome and targeted sequencing data done in this paper have been deposited at the European Genome-phenome Archive (EGA; http://www.ebi.ac.uk/ega) under accession numbers EGAS00001001653 and at the International Cancer Genome Consortium Data Portal database (https://dcc.icgc.org/). The whole exome sequencing data (8) has been deposited at EGA under accession number EGA00001000950. Gene expression and Methylation data have been deposited at the Gene Expression Omnibus (GEO; www.ncbi.nlm.nih.gov/geo) with GEO accession numbers GSE89749 and GSE89803, respectively.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Apinya Jusakul[1,2,3,*], Ioana Cutcutache[4,*], Chern Han Yong[1,4,*], Jing Quan Lim[2,5,*], Mi Ni Huang[4], Nisha Padmanabhan[1], Vishwa Nellore[6], Sarinya Kongpetch[2,7,8], Alvin Wei Tian Ng[9], Ley Moy Ng[10], Su Pin Choo[11], Swe Swe Myint[2], Raynoo Thanan[12], Sanjanaa Nagarajan[2], Weng Khong Lim[1,2], Cedric Chuan Young Ng[2], Arnoud Boot[1,4], Mo Liu[1,4], Choon Kiat Ong[5], Vikneswari Rajasegaran[2], Stefanus Lie[2,13], Alvin Soon Tiong Lim[14], Tse Hui Lim[14], Jing Tan[2], Jia Liang Loh[2], John R. McPherson[4], Narong Khuntikeo[7,15], Vajaraphongsa Bhudhisawasdi[15], Puangrat

Yongvanit[7], Sopit Wongkham[12], Yasushi Totoki[16], Hiromi Nakamura[16], Yasuhito Arai[16], Satoshi Yamasaki[17], Pierce Kah-Hoe Chow[18], Alexander Yaw Fui Chung[19], London Lucien Peng Jin Ooi[19], Kiat Hon Lim[20], Simona Dima[21], Dan G. Duda[22], Irinel Popescu[21], Philippe Broet[23], Sen-Yung Hsieh[24], Ming-Chin Yu[25], Aldo Scarpa[26], Jiaming Lai[27], Di-Xian Luo[28], André Lopes Carvalho[29], André Luiz Vettore[30], Hyungjin Rhee[31], Young Nyun Park[31], Ludmil B. Alexandrov[32], Raluca Gordân[6,33], Steven G. Rozen[1,4,34], Tatsuhiro Shibata[16,17], Chawalit Pairojkul[35], Bin Tean Teh[1,2,10,34,36], and Patrick Tan[1,10,34,37]

Affiliations

[1]Program in Cancer and Stem Cell Biology, Duke-NUS Medical School, Singapore [2]Laboratory of Cancer Epigenome, Division of Medical Science, National Cancer Centre Singapore, Singapore [3]The Centre for Research and Development of Medical Diagnostic Laboratories and Department of Clinical Immunology and Transfusion Sciences., Faculty of Associated Medical Sciences, Khon Kaen University, Khon Kaen, Thailand [4]Centre for Computational Biology, Duke-NUS Medical School, Singapore [5]Lymphoma Genomic Translational Research Laboratory, National Cancer Centre Singapore, Division of Medical Oncology, Singapore [6]Department of Biostatistics and Bioinformatics, Center for Genomic and Computational Biology, Duke University, Durham, North Carolina, USA [7]Cholangiocarcinoma Screening and Care Program and Liver Fluke and Cholangiocarcinoma Research Centre, Faculty of Medicine, Khon Kaen University, Khon Kaen, Thailand [8]Department of Pharmacology, Faculty of Medicine, Khon Kaen University, Khon Kaen, Thailand [9]NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore [10]Cancer Science Institute of Singapore, National University of Singapore, Singapore [11]Division of Medical Oncology, National Cancer Centre Singapore, Singapore [12]Department of Biochemistry, Faculty of Medicine, Khon Kaen University, Khon Kaen, Thailand [13]Division of Radiation Oncology, National Cancer Centre Singapore, Singapore [14]Cytogenetics Laboratory, Department of Molecular Pathology, Singapore General Hospital, Singapore [15]Department of Surgery, Faculty of Medicine, Khon Kaen University, Khon Kaen, Thailand [16]Division of Cancer Genomics, National Cancer Center Research Institute, Tokyo, Japan [17]Laboratory of Laboratory of Molecular Medicine, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Japan [18]Division of Surgical Oncology, National Cancer Center Singapore and Office of Clinical Sciences, Duke-NUS Medical School, Singapore [19]Department of Hepatopancreatobiliary/Transplant Surgery, Singapore General Hospital, Singapore [20]Department of Anatomical Pathology, Singapore General Hospital, Singapore [21]Center of Digestive Diseases and Liver Transplantation, Fundeni Clinical Institute, Bucharest, Romania [22]Edwin L. Steele Laboratories for Tumor Biology, Department of Radiation Oncology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA [23]DHU Hepatinov, Hôpital Paul Brousse, AP-HP, Villejuif, France [24]Department of Gastroenterology and Hepatology, Chang Gung Memorial Hospital and Chang Gung University, Taoyuan 333, Taiwan [25]Department of General Surgery, Chang

Gung Memorial Hospital and Chang Gung University, Taoyuan 333, Taiwan [26]Applied Research on Cancer Centre (ARC-Net), University and Hospital Trust of Verona, Verona, Italy [27]Department of Hepatobiliary Surgery, the First Affiliated Hospital of Sun Yat-sen University, Guangzhou, P.R. China [28]National and Local Joint Engineering Laboratory of High-through Molecular Diagnostic Technology, the First People's Hospital of Chenzhou, Southern Medical University, Chenzhou, P.R. China [29]Barretos Cancer Hospital, Barretos, SP, Brazil [30]Laboratory of Cancer Molecular Biology, Department of Biological Sciences, Federal University of São Paulo, Rua Pedro de Toledo 669, SP, Brazil [31]Department of Pathology, Brain Korea 21 PLUS Project for Medical Science, Integrated Genomic Research Center for Metabolic Regulation, Yonsei University College of Medicine, Seoul, Korea [32]Theoretical Biology and Biophysics (T-6), Los Alamos National Laboratory, and Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico, USA [33]Department of Computer Science, Duke University, Durham, North Carolina, USA [34]SingHealth/Duke-NUS Institute of Precision Medicine, National Heart Centre, Singapore [35]Department of Pathology, Faculty of Medicine, Khon Kaen University, Khon Kaen, Thailand [36]Institute of Molecular and Cell Biology, Singapore [37]Genome Institute of Singapore, Singapore

Acknowledgments

This study is performed as part of the International Cancer Genome Consortium (ICGC). We thank the Duke–NUS Genome Biology Facility for methylation and gene expression assays, the GIS Population Genetics facility for SNP arrays and the Cytogenetics Laboratory, Department of Molecular Pathology, Singapore General Hospital for FISH analysis. We thank D. Jefferson (New England Medical Center, Tufts University) for H69 cells. We also thank the SingHealth Tissue Repository, Dr. Catherine Guettier and Dr. Jean-Charles Duclos-Vallée (DHU Hepatinov, Hôpital Paul Brousse, AP-HP, Villejuif, France) for tissue samples.

Financial support:

We thank Sir Lambert Cornelias Bronsveld and Lady Bronsveld-Ngo Kim Lian for philanthropic support (B.T.T.). This work was supported by the Singapore National Medical Research Council (NMRC/STaR/0006/2009 (B.T.T.), NMRC/STaR/0024/2014 (B.T.T.), NMRC/CG/012/2013 (B.T.T.), NMRC/CIRG/1422/2015 (S.R.), and NMRC/STaR/0026/2015 (P.T.)), Genome Institute of Singapore (P.T.), Duke-NUS Medical School (P.T., S.R.), National University of Singapore, the National Research Foundation Singapore, Singapore Ministry of Education under the Research Centres of Excellence initiative (P.T), Japan Agency for Medical Research and Development (Practical Research for Innovative Cancer Control, 15ck0106094h0002) (T.S.), National Cancer Center Research and Development Funds (26-A-5) (T.S.), Italian Cancer Genome Project (FIRB RBAP10AHJB) (A.S.), Associazione Italiana Ricerca sul Cancro (n. 12182) (A.S.), Italian Ministry of University Health (FIMP CUP_J33G13000210001) (A.S.), National Cancer Institute of the National Institutes of Health of the United States of America (P01CA142538) (R.G.) and National Natural Science Foundation of China (NSFC81372825) (J.L.).

References

1. Banales JM, Cardinale V, Carpino G, Marzioni M, Andersen JB, Invernizzi P, et al. Expert consensus document: Cholangiocarcinoma: current knowledge and future perspectives consensus statement from the European Network for the Study of Cholangiocarcinoma (ENS-CCA). Nat Rev Gastroenterol Hepatol. 2016; 13:261–80. [PubMed: 27095655]
2. Jusakul A, Kongpetch S, Teh BT. Genetics of Opisthorchis viverrini-related cholangiocarcinoma. Curr Opin Gastroenterol. 2015; 31:258–63. [PubMed: 25693006]

3. Kongpetch S, Jusakul A, Ong CK, Lim WK, Rozen SG, Tan P, et al. Pathogenesis of cholangiocarcinoma: From genetics to signalling pathways. Best Pract Res Clin Gastroenterol. 2015; 29:233–44. [PubMed: 25966424]

4. Khuntikeo N, Chamadol N, Yongvanit P, Loilome W, Namwat N, Sithithaworn P, et al. Cohort profile: cholangiocarcinoma screening and care program (CASCAP). BMC Cancer. 2015; 15:459. [PubMed: 26054405]

5. Chen JS, Hsu C, Chiang NJ, Tsai CS, Tsou HH, Huang SF, et al. A KRAS mutation status-stratified randomized phase II trial of gemcitabine and oxaliplatin alone or in combination with cetuximab in advanced biliary tract cancer. Ann Oncol. 2015; 26:943–9. [PubMed: 25632066]

6. Ong CK, Subimerb C, Pairojkul C, Wongkham S, Cutcutache I, Yu W, et al. Exome sequencing of liver fluke-associated cholangiocarcinoma. Nat Genet. 2012; 44:690–3. [PubMed: 22561520]

7. Chan-On W, Nairismagi ML, Ong CK, Lim WK, Dima S, Pairojkul C, et al. Exome sequencing identifies distinct mutational patterns in liver fluke-related and non-infection-related bile duct cancers. Nat Genet. 2013; 45:1474–8. [PubMed: 24185513]

8. Nakamura H, Arai Y, Totoki Y, Shirota T, Elzawahry A, Kato M, et al. Genomic spectra of biliary tract cancer. Nat Genet. 2015; 47:1003–10. [PubMed: 26258846]

9. Waddell N, Pajic M, Patch AM, Chang DK, Kassahn KS, Bailey P, et al. Whole genomes redefine the mutational landscape of pancreatic cancer. Nature. 2015; 518:495–501. [PubMed: 25719666]

10. Fujimoto A, Furuta M, Shiraishi Y, Gotoh K, Kawakami Y, Arihiro K, et al. Whole-genome mutational landscape of liver cancers displaying biliary phenotype reveals hepatitis impact and molecular diversity. Nat Commun. 2015; 6:6120. [PubMed: 25636086]

11. Fujimoto A, Totoki Y, Abe T, Boroevich KA, Hosoda F, Nguyen HH, et al. Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. Nat Genet. 2012; 44:760–4. [PubMed: 22634756]

12. Tyson GL, El-Serag HB. Risk factors for cholangiocarcinoma. Hepatology. 2011; 54:173–84. [PubMed: 21488076]

13. Farshidfar F, Zheng S, Gingras MC, Newton Y, Shih J, Robertson AG, et al. Integrative Genomic Analysis of Cholangiocarcinoma Identifies Distinct IDH-Mutant Molecular Profiles. Cell reports. 2017; 18:2780–94. [PubMed: 28297679]

14. Jiao Y, Pawlik TM, Anders RA, Selaru FM, Streppel MM, Lucas DJ, et al. Exome sequencing identifies frequent inactivating mutations in BAP1, ARID1A and PBRM1 in intrahepatic cholangiocarcinomas. Nat Genet. 2013; 45:1470–3. [PubMed: 24185509]

15. Zou S, Li J, Zhou H, Frech C, Jiang X, Chu JSC, et al. Mutational landscape of intrahepatic cholangiocarcinoma. Nature Communications. 2014; 5:5696.

16. Simbolo M, Fassan M, Ruzzenente A, Mafficini A, Wood LD, Corbo V, et al. Multigene mutational profiling of cholangiocarcinomas identifies actionable molecular subgroups. Oncotarget. 2014; 5:2839–52. [PubMed: 24867389]

17. Lee H, Wang K, Johnson A, Jones DM, Ali SM, Elvin JA, et al. Comprehensive genomic profiling of extrahepatic cholangiocarcinoma reveals a long tail of therapeutic targets. J Clin Pathol. 2016; 69:403–8. [PubMed: 26500333]

18. Harbour JW, Roberson ED, Anbunathan H, Onken MD, Worley LA, Bowcock AM. Recurrent mutations at codon 625 of the splicing factor SF3B1 in uveal melanoma. Nat Genet. 2013; 45:133–5. [PubMed: 23313955]

19. Maguire SL, Leonidou A, Wai P, Marchio C, Ng CK, Sapino A, et al. SF3B1 mutations constitute a novel therapeutic target in breast cancer. J Pathol. 2015; 235:571–80. [PubMed: 25424858]

20. Treekitkarnmongkol W, Suthiphongchai T. High expression of ErbB2 contributes to cholangiocarcinoma cell invasion and proliferation through AKT/p70S6K. World J Gastroenterol. 2010; 16:4047–54. [PubMed: 20731018]

21. Arai Y, Totoki Y, Hosoda F, Shirota T, Hama N, Nakamura H, et al. Fibroblast growth factor receptor 2 tyrosine kinase fusions define a unique molecular subtype of cholangiocarcinoma. Hepatology. 2014; 59:1427–34. [PubMed: 24122810]

22. Wu YM, Su F, Kalyana-Sundaram S, Khazanov N, Ateeq B, Cao X, et al. Identification of targetable FGFR gene fusions in diverse cancers. Cancer Discov. 2013; 3:636–47. [PubMed: 23558953]

23. Capelletti M, Dodge ME, Ercan D, Hammerman PS, Park SI, Kim J, et al. Identification of recurrent FGFR3-TACC3 fusion oncogenes from lung adenocarcinoma. Clin Cancer Res. 2014; 20:6551–8. [PubMed: 25294908]
24. Williams SV, Hurst CD, Knowles MA. Oncogenic FGFR3 gene fusions in bladder cancer. Hum Mol Genet. 2013; 22:795–803. [PubMed: 23175443]
25. Kataoka K, Shiraishi Y, Takeda Y, Sakata S, Matsumoto M, Nagano S, et al. Aberrant PD-L1 expression through 3′-UTR disruption in multiple cancers. Nature. 2016; 534:402–6. [PubMed: 27281199]
26. Byron SA, Chen H, Wortmann A, Loch D, Gartside MG, Dehkhoda F, et al. The N550K/H mutations in FGFR2 confer differential resistance to PD173074, dovitinib, and ponatinib ATP-competitive inhibitors. Neoplasia. 2013; 15:975–88. [PubMed: 23908597]
27. Chen P, Zhang L, Weng T, Zhang S, Sun S, Chang M, et al. A Ser252Trp mutation in fibroblast growth factor receptor 2 (FGFR2) mimicking human Apert syndrome reveals an essential role for FGF signaling in the regulation of endochondral bone formation. PLoS One. 2014; 9:e87311. [PubMed: 24489893]
28. Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ 3rd, et al. Landscape of somatic retrotransposition in human cancers. Science. 2012; 337:967–71. [PubMed: 22745252]
29. Iskow RC, McCabe MT, Mills RE, Torene S, Pittard WS, Neuwald AF, et al. Natural mutagenesis of human genomes by endogenous retrotransposons. Cell. 2010; 141:1253–61. [PubMed: 20603005]
30. Kazazian HH Jr, Goodier JL. LINE drive retrotransposition and genome instability. Cell. 2002; 110:277–80. [PubMed: 12176313]
31. Weinhold N, Jacobsen A, Schultz N, Sander C, Lee W. Genome-wide analysis of noncoding regulatory mutations in cancer. Nat Genet. 2014; 46:1160–5. [PubMed: 25261935]
32. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, et al. Integrative annotation of variants from 1092 humans: application to cancer genomics. Science. 2013; 342:1235587. [PubMed: 24092746]
33. Melton C, Reuter JA, Spacek DV, Snyder M. Recurrent somatic mutations in regulatory regions of human cancer genomes. Nat Genet. 2015; 47:710–6. [PubMed: 26053494]
34. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 2014; 158:1431–43. [PubMed: 25215497]
35. Berger MF, Bulyk ML. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nat Protoc. 2009; 4:393–411. [PubMed: 19265799]
36. Siggers T, Gordan R. Protein-DNA binding: complexities and multi-protein codes. Nucleic Acids Res. 2014; 42:2099–111. [PubMed: 24243859]
37. Efron B, Tibshirani R. On testing the significance of sets of genes. 2007:107–29.
38. Ichimura N, Shinjo K, An B, Shimizu Y, Yamao K, Ohka F, et al. Aberrant TET1 Methylation Closely Associated with CpG Island Methylator Phenotype in Colorectal Cancer. Cancer Prev Res (Phila). 2015; 8:702–11. [PubMed: 26063725]
39. Wang P, Dong Q, Zhang C, Kuan PF, Liu Y, Jeck WR, et al. Mutations in isocitrate dehydrogenase 1 and 2 occur frequently in intrahepatic cholangiocarcinomas and share hypermethylation targets with glioblastomas. Oncogene. 2013; 32:3091–100. [PubMed: 22824796]
40. Sato Y, Yoshizato T, Shiraishi Y, Maekawa S, Okuno Y, Kamura T, et al. Integrated molecular analysis of clear-cell renal cell carcinoma. Nat Genet. 2013; 45:860–7. [PubMed: 23797736]
41. Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S, et al. Clock-like mutational processes in human somatic cells. Nat Genet. 2015; 47:1402–7. [PubMed: 26551669]
42. Cooper DN, Mort M, Stenson PD, Ball EV, Chuzhanova NA. Methylation-mediated deamination of 5-methylcytosine appears to give rise to mutations causing human inherited disease in CpNpG trinucleotides, as well as in CpG dinucleotides. Hum Genomics. 2010; 4:406–10. [PubMed: 20846930]
43. Chan E, Berlin J. Biliary tract cancers: understudied and poorly understood. J Clin Oncol. 2015; 33:1845–8. [PubMed: 25918294]

44. Malka D, Cervera P, Foulon S, Trarbach T, de la Fouchardiere C, Boucher E, et al. Gemcitabine and oxaliplatin with or without cetuximab in advanced biliary-tract cancer (BINGO): a randomised, open-label, non-comparative phase 2 trial. Lancet Oncol. 2014; 15:819–28. [PubMed: 24852116]
45. Law LY. Dramatic response to trastuzumab and paclitaxel in a patient with human epidermal growth factor receptor 2-positive metastatic cholangiocarcinoma. J Clin Oncol. 2012; 30:e271–3. [PubMed: 22851567]
46. Senovilla L, Vitale I, Martins I, Tailler M, Pailleret C, Michaud M, et al. An immunosurveillance mechanism controls cancer cell ploidy. Science. 2012; 337:1678–84. [PubMed: 23019653]
47. Durrbaum M, Kuznetsova AY, Passerini V, Stingele S, Stoehr G, Storchova Z. Unique features of the transcriptional response to model aneuploidy in human cells. BMC Genomics. 2014; 15:139. [PubMed: 24548329]
48. Schottlandt, D. ArQule Announces Orphan Drug Designation in Cholangiocarcinoma and Clinical Update for ARQ 087. 2015. Acquire Media <http://investors.arqule.com/releasedetail.cfm?ReleaseID=947001>
49. Fredriksson NJ, Ny L, Nilsson JA, Larsson E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. Nat Genet. 2014; 46:1258–63. [PubMed: 25383969]
50. Young ND, Nagarajan N, Lin SJ, Korhonen PK, Jex AR, Hall RS, et al. The Opisthorchis viverrini genome provides insights into life in the bile duct. Nat Commun. 2014; 5:4378. [PubMed: 25007141]
51. Chaiyadet S, Sotillo J, Smout M, Cantacessi C, Jones MK, Johnson MS, et al. Carcinogenic Liver Fluke Secretes Extracellular Vesicles That Promote Cholangiocytes to Adopt a Tumorigenic Phenotype. J Infect Dis. 2015; 212:1636–45. [PubMed: 25985904]
52. Brandi G, Farioli A, Astolfi A, Biasco G, Tavolari S. Genetic heterogeneity in cholangiocarcinoma: a major challenge for targeted therapies. Oncotarget. 2015; 6:14744–53. [PubMed: 26142706]
53. Grubman SA, Perrone RD, Lee DW, Murray SL, Rogers LC, Wolkoff LI, et al. Regulation of intracellular pH by immortalized human intrahepatic biliary epithelial cell lines. Am J Physiol. 1994; 266:G1060–70. [PubMed: 8023938]
54. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010; 26:589–95. [PubMed: 20080505]
55. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–9. [PubMed: 19505943]
56. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20:1297–303. [PubMed: 20644199]
57. Gordan R, Murphy KF, McCord RP, Zhu C, Vedenko A, Bulyk ML. Curated collection of yeast transcription factor DNA binding specificity data reveals novel structural and gene regulatory insights. Genome Biol. 2011; 12:R125. [PubMed: 22189060]
58. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013; 499:214–8. [PubMed: 23770567]
59. Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, et al. IntOGen-mutations identifies cancer drivers across tumor types. Nat Methods. 2013; 10:1081–2. [PubMed: 24037244]
60. Wang J, Mullighan CG, Easton J, Roberts S, Heatley SL, Ma J, et al. CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nat Methods. 2011; 8:652–4. [PubMed: 21666668]
61. Collins JR, Stephens RM, Gold B, Long B, Dean M, Burt SK. An exhaustive DNA micro-satellite map of the human genome using high performance computing. Genomics. 2003; 82:10–9. [PubMed: 12809672]
62. Wang J, Song L, Grover D, Azrak S, Batzer MA, Liang P. dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. Hum Mutat. 2006; 27:323–9. [PubMed: 16511833]

63. Van Loo P, Nordgard SH, Lingjaerde OC, Russnes HG, Rye IH, Sun W, et al. Allele-specific copy number analysis of tumors. Proc Natl Acad Sci U S A. 2010; 107:16910–5. [PubMed: 20837533]
64. Nilsen G, Liestøl K, Van Loo P, Moen Vollan HK, Eide MB, Rueda OM, et al. Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. BMC Genomics. 2012; 13:591. [PubMed: 23442169]
65. Boeva V, Popova T, Bleakley K, Chiche P, Cappo J, Schleiermacher G, et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. Bioinformatics. 2012; 28:423–5. [PubMed: 22155870]
66. Reinecke F, Satya RV, DiCarlo J. Quantitative analysis of differences in copy numbers using read depth obtained from PCR-enriched samples and controls. BMC Bioinformatics. 2015; 16:17. [PubMed: 25626454]
67. Favero F, Joshi T, Marquard AM, Birkbak NJ, Krzystanek M, Li Q, et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. Ann Oncol. 2015; 26:64–70. [PubMed: 25319062]
68. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011; 12:R41. [PubMed: 21527027]
69. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. Bioinformatics. 2008; 24:1547–8. [PubMed: 18467348]
70. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007; 8:118–27. [PubMed: 16632515]
71. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005; 102:15545–50.
72. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nature Communications. 2013; 4:2612.
73. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014; 30:1363–9. [PubMed: 24478339]
74. Pidsley R, Y Wong CC, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. BMC Genomics. 2013; 14:293. [PubMed: 23631413]
75. Koestler DC, Christensen BC, Marsit CJ, Kelsey KT, Houseman EA. Recursively partitioned mixture model clustering of DNA methylation data using biologically informed correlation structures. Stat Appl Genet Mol Biol. 2013; 12:225–40. [PubMed: 23468465]
76. Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc Natl Acad Sci U S A. 2013; 110:4245–50. [PubMed: 23431203]

STATEMENT OF SIGNIFICANCE

Integrated whole-genome and epigenomic analysis of cholangiocarcinoma (CCA) on an international scale identifies new CCA driver genes, non-coding promoter mutations, and structural variants. CCA molecular landscapes differ radically by etiology, underscoring how distinct cancer subtypes in the same organ may arise through different extrinsic and intrinsic carcinogenic processes.

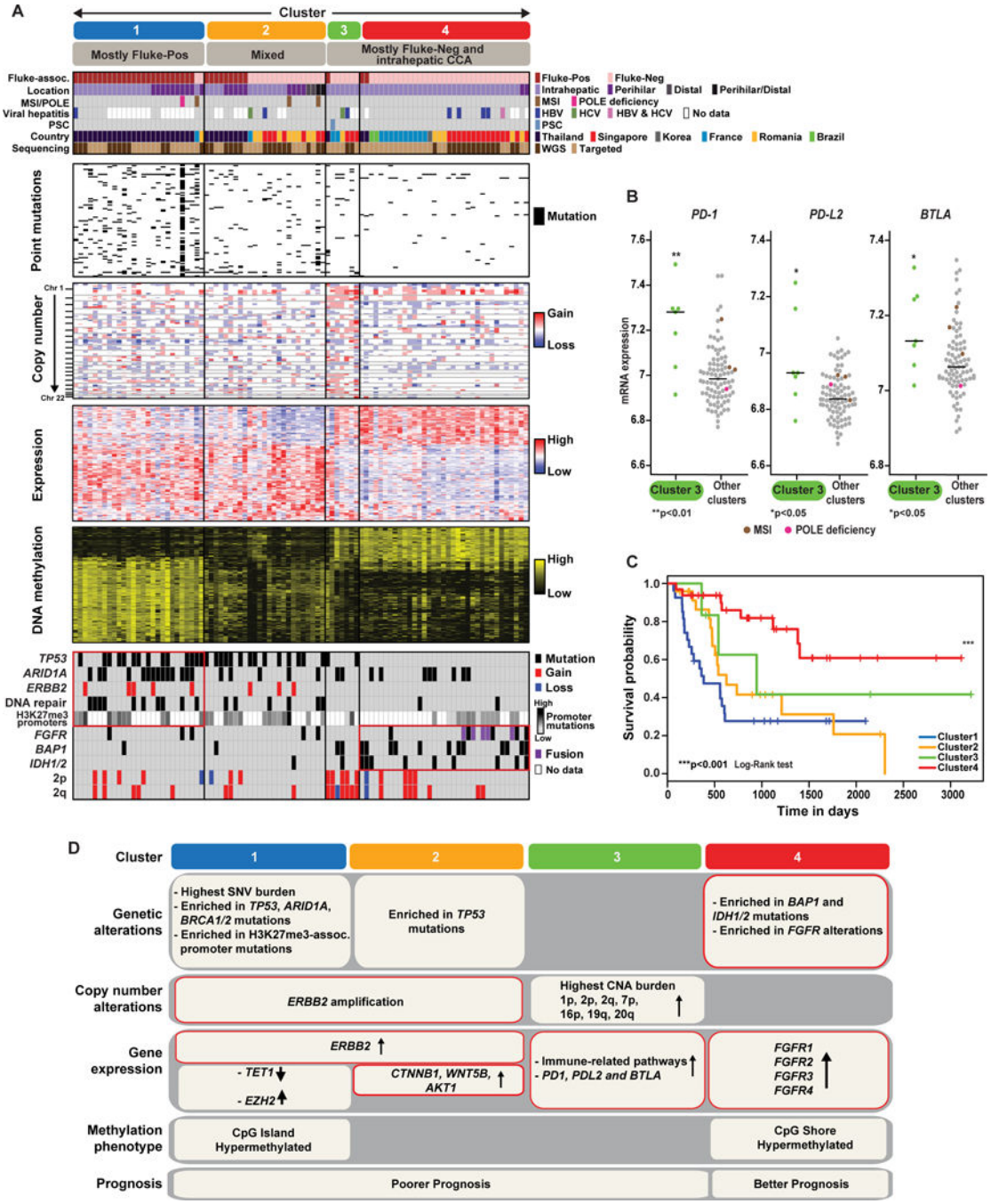

Figure 1. Integrative Clustering Defines Four Molecular Subtypes of CCA

(A) Heatmap showing four clusters identified by iClusterPlus based on clustering of mutation, copy-number, gene expression and methylation data. Top rows indicate clinical characteristics, risk factors, geographical region, and sequencing platform. Microsatellite instability (MSI) status was defined by indel counts (≥ 6 indels) in simple repeat sequences. Bottom rows indicate selected genetic alterations.

(B) High expression of *PD-1, PD-L2* and *BTLA* in Cluster 3 relative to other clusters. Brown dots indicate MSI cases. Pink dots indicate cases with DNA polymerase epsilon (*POLE*) proofreading deficiency.

(C) Survival analysis showing improved survival in Cluster 3 and 4 CCAs compared to other clusters. Multivariate analysis confirmed this difference even after accounting for fluke association, anatomical location, and clinical staging.

(D) Representative genetic, epigenetic and gene expression features of CCA clusters.

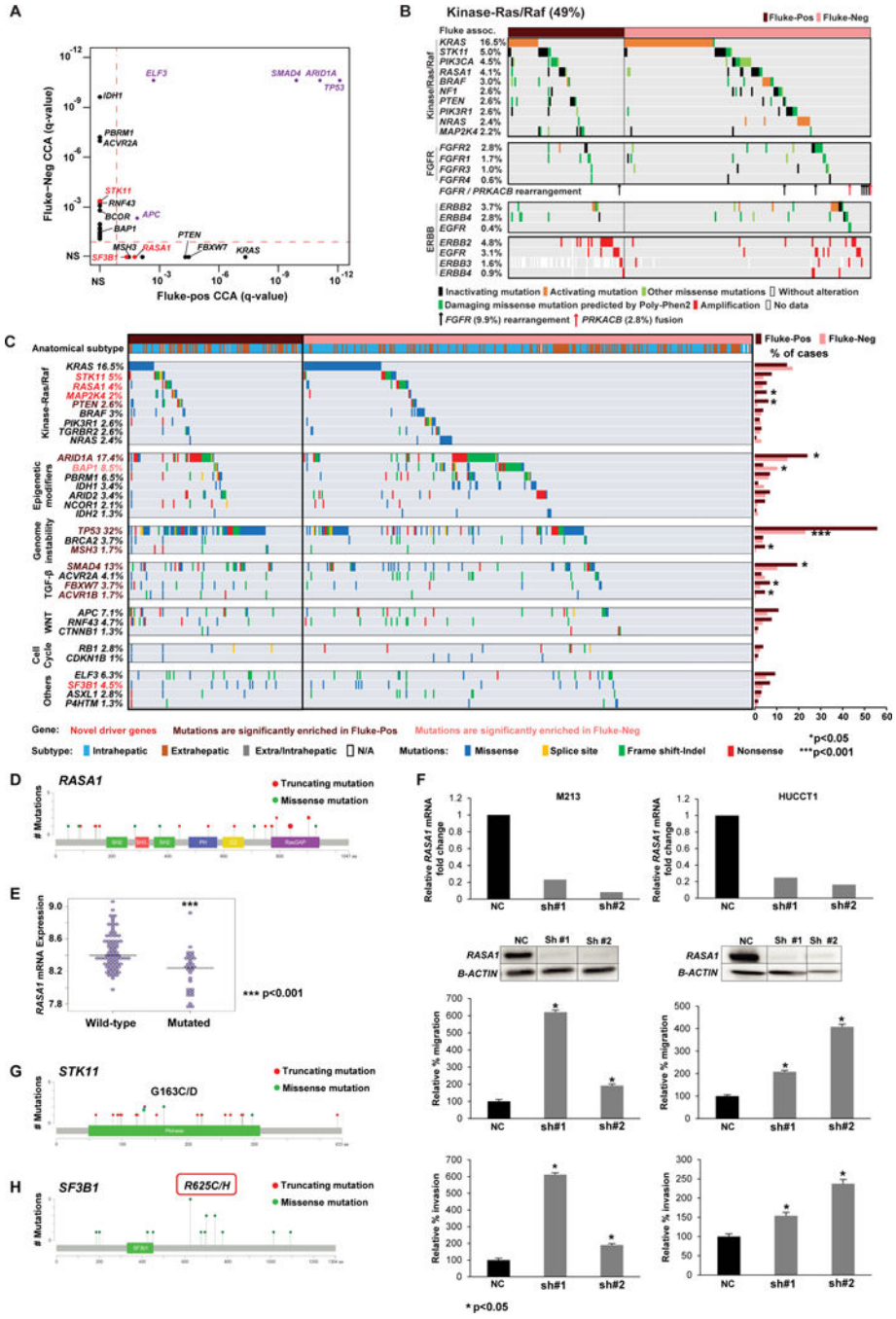

Figure 2. Significantly Mutated Genes in CCAs

(A) Significantly mutated genes in Fluke-Pos and Fluke-Neg CCA. Genes in purple are mutated in both Fluke-Pos and Fluke-Neg CCAs. Novel significantly mutated genes are highlighted in red.

(B) Alterations in Kinase-Ras/Raf pathway components across Fluke-Pos and Fluke-Neg CCAs. CCAs with *FGFR/PRKACB* rearrangements are also highlighted (arrows).

(C) Matrix of genes (rows) and tumors (columns) showing occurrence of 32 somatic mutated genes. The bar chart at right shows frequencies of affected cases in Fluke-Pos and Fluke-

Neg tumors. Asterisks indicate genes with significant differences between Fluke-Pos and Fluke-Neg CCAs. P-values were computed using the Fisher's exact test.
(D) Distribution of somatic mutations in *RASA1*.
(E) *RASA1* expression in tumors without *RASA1* alterations (Wild-type) compared to tumors with *RASA1* deletions and inactivating mutations (nonsense mutations or frame-shift indels).
(F) *RASA1* shRNA silencing inhibits CCA migration and invasion *in vitro*. Expression levels (mRNA and protein) of *RASA1* in M213 (left) and HUCCT1 (right) cells transduced with two independent shRNAs (*RASA1* shRNA#1 and *RASA1* shRNA#2) targeting different regions of *RASA1* were assessed by qPCR and Western blotting (first and second panel, respectively). Migration and invasion of *RASA1* knockdown cells were assessed by transwell assays. Mean ± SEM of three independent experiments were analyzed.
(G–H) Distribution of somatic mutations in *STK11* (G) and *SF3B1* (H). The red box indicates mutations in previously described *SF3B1* hotspots.

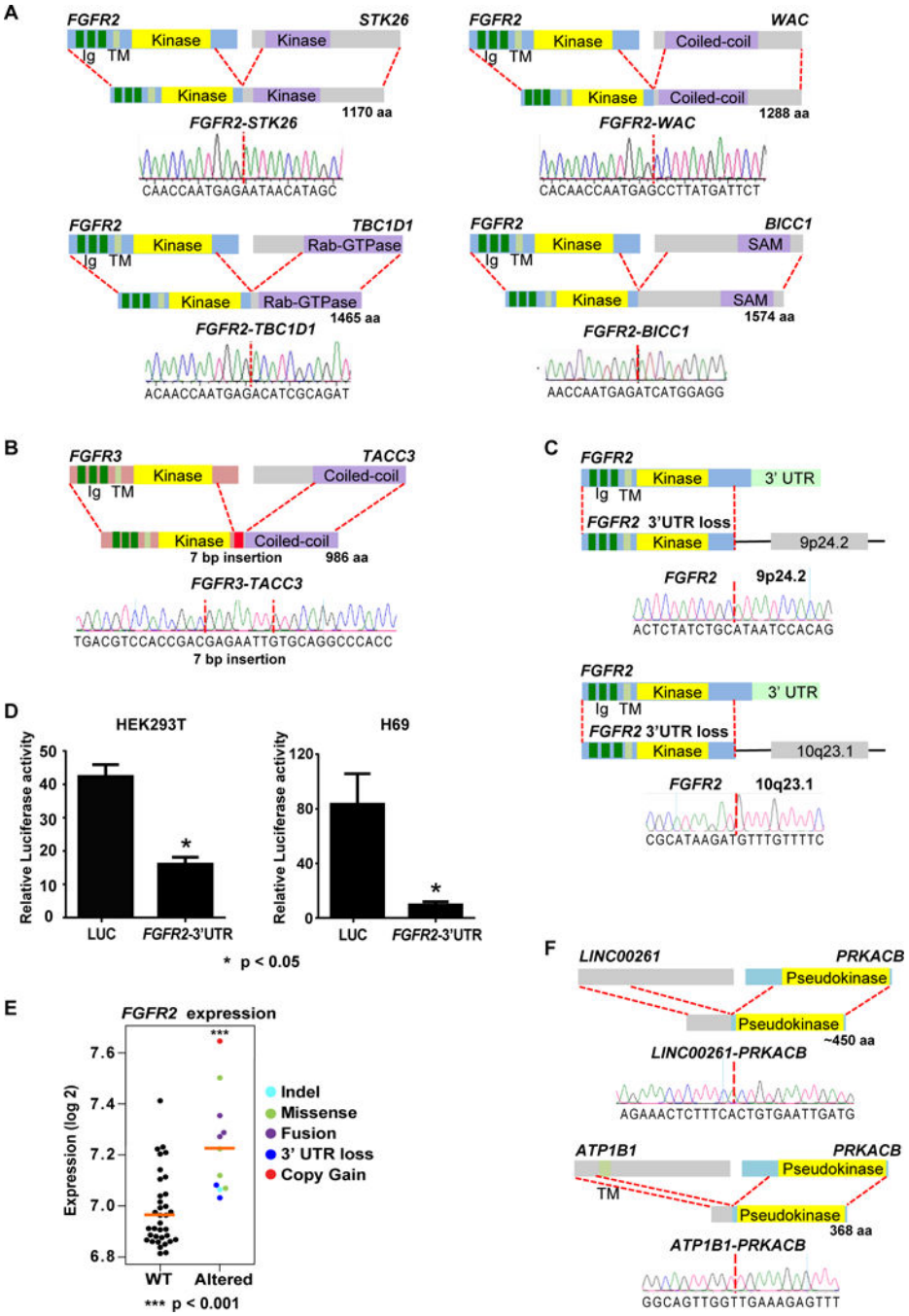

Figure 3. *FGFR* and *PRKA* Gene Rearrangements in CCA

(A) Identification of *FGFR2-STK26*, *FGFR2-WAC* and *FGFR2-TBC1D1* and *FGFR2-BICC1* rearrangements in CCA. All fusions were validated by RT-PCR and sequence chromatograms are shown. *FGFR2-STK26*, *FGFR2-WAC* and *FGFR2-TBC1D1* were validated in this study, while *FGFR2-BICC1* was validated in (8).

(B) Identification of a *FGFR3-TACC3* gene fusion. Transcript validation was performed confirming a 7 bp insertion (red dotted lines).

(C) Recurrent loss of 3′ UTRs in *FGFR2* due to rearrangements with intergenic regions.

(D) Relative luciferase activity between empty luciferase vector (LUC) and *FGFR2*-3′UTR in HEK293T and H69 immortalized cholangiocyte cell lines. Data is presented in Mean ± SD. Three individual experiments were performed.
(E) *FGFR2* gene expression levels between *FGFR2*-wildtype CCAs and CCAs exhibiting different categories of *FGFR2* alterations, as shown by the color chart.
(F) Identification of *LINC00261-PRKACB* and *ATP1B1-PRKACB* fusions. Both fusions were validated by RT-PCR and sequence chromatograms.

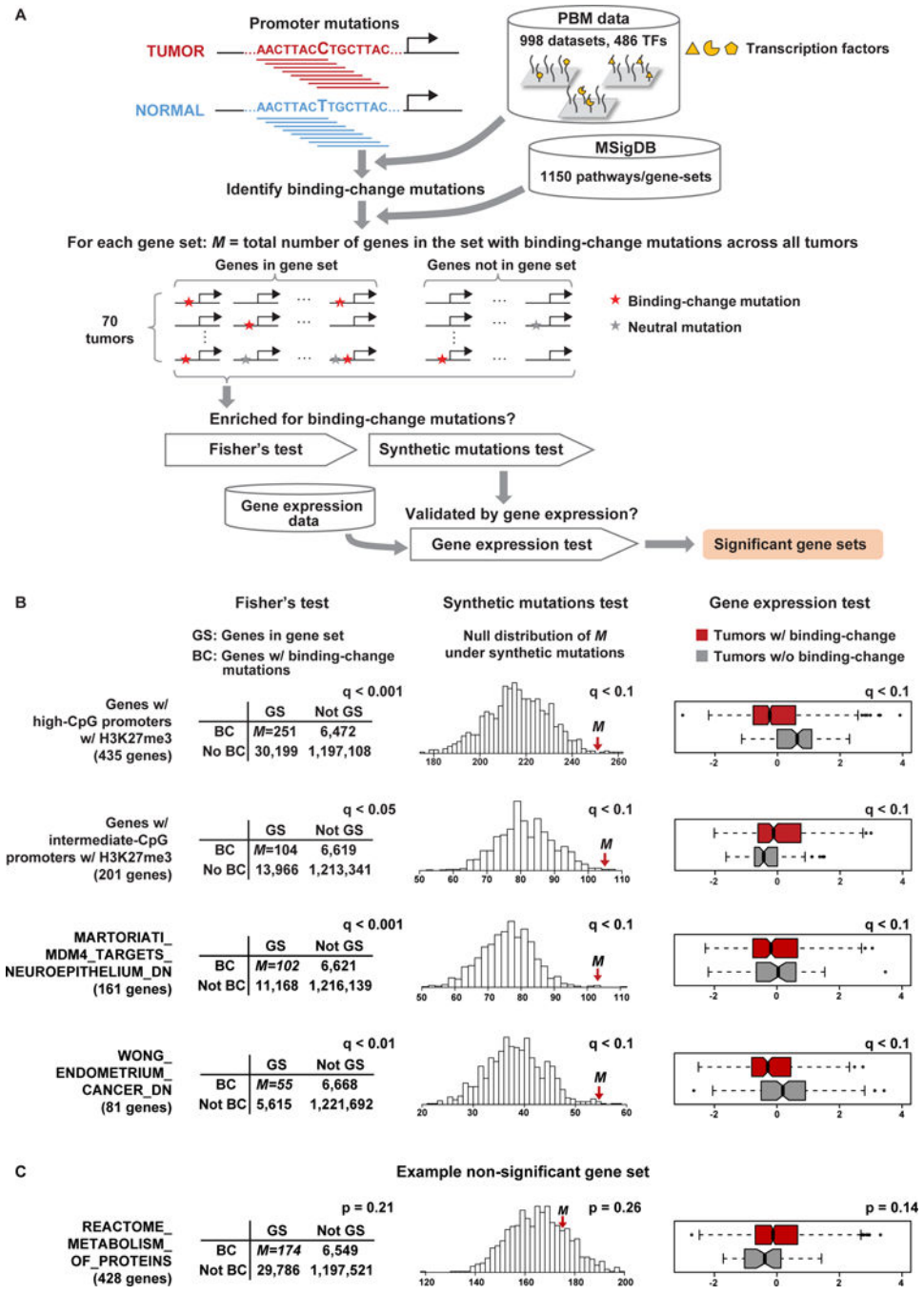

Figure 4. FIREFLY Analysis of Pathways Systematically Dysregulated by Somatic Promoter Mutations that Alter Transcription Factor Binding

(A) Changes in transcription factor (TF)-DNA binding estimated from PBM (protein binding microarray) data for 486 mammalian TFs. Changes in binding specificity were computed using PBM-derived binding scores for 8-mer sequences overlapping each mutation. To determine whether a given gene set was preferentially enriched for binding-change mutations, we computed the statistic *M* (the number of genes in the gene set with TF binding-change mutations in the promoter) summed over all tumors. FIREFLY assessed systematic enrichment of binding-change mutations with 2 statistical tests: (i) Fisher's exact

test of whether *M* is greater than expected by chance given the number genes in the gene set and the total number of genes affected by binding-change mutations, (ii) a comparison of *M* in actual data to a null distribution of *M* over 1,000 sets of 70 *in-silico* mutated tumor sequences, based on patient-specific trinucleotide contexts of mutations for each tumor. FIREFLY then tests for putative transcriptional dysregulation associated with the binding-change mutations by performing a Gene Set Analysis (GSA) to associate gene expression dysregulation with the number of binding-change mutations.

(B) Details of the four gene sets meeting FIREFLY's criteria of $q < 0.1$ for all three statistical tests: MIKKELSEN_MCV6_HCP_WITH_H3K27ME3, MIKKELSEN_MEF_ICP_WITH_H3K27ME3, MARTORIATI_MDM4_TARGETS_NEUROEPITHELIUM_DN, and WONG_ENDOMETRIUM_CANCER_DN.

(C) Details of an example non-significant gene set.

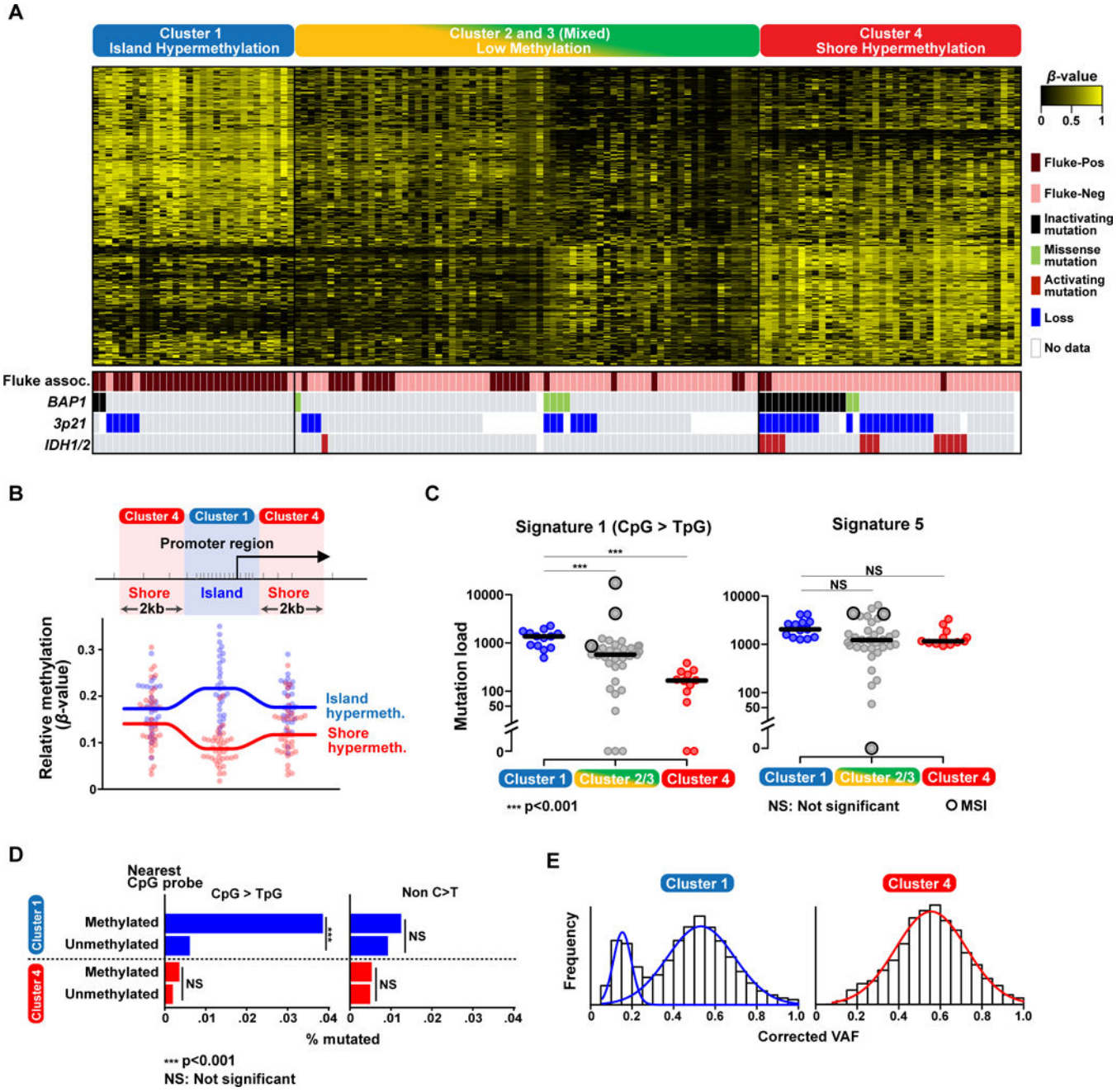

Figure 5. Epigenetic Clusters and Integration of Mutation Signatures in CCA

(A) Heatmap showing three DNA methylation clusters: two hypermethylated clusters (Cluster 1 and Cluster 4), and a low methylation cluster (mixed Clusters 2 and 3).

(B) Distinct methylation patterns in Cluster 1 versus Cluster 4. Top: typical organization of a gene promoter with CpG island and shores. Vertical ticks represent CpG sites. Bottom: Levels of promoter hypermethylation in CpG islands and shores in Cluster 1 and Cluster 4.

(C) Left: enrichment of mutation Signature 1 (CpG>TpG) in Cluster 1. Right: similar levels of mutation Signature 5 among methylation clusters. Circled tumors represent MSI tumors.

(D) Proximity of somatic mutations to hypermethylated CpGs in Cluster 1 and 4. Left: CpG>TpG mutations are located preferentially near hypermethylated CpGs in Cluster 1, but not in Cluster 4. Right: Non C>T mutations are not located preferentially near hypermethylated CpGs in either Cluster 1 or 4.
(E) Histograms of corrected variant allele frequencies (VAF) of point mutations in Cluster 1 and 4.

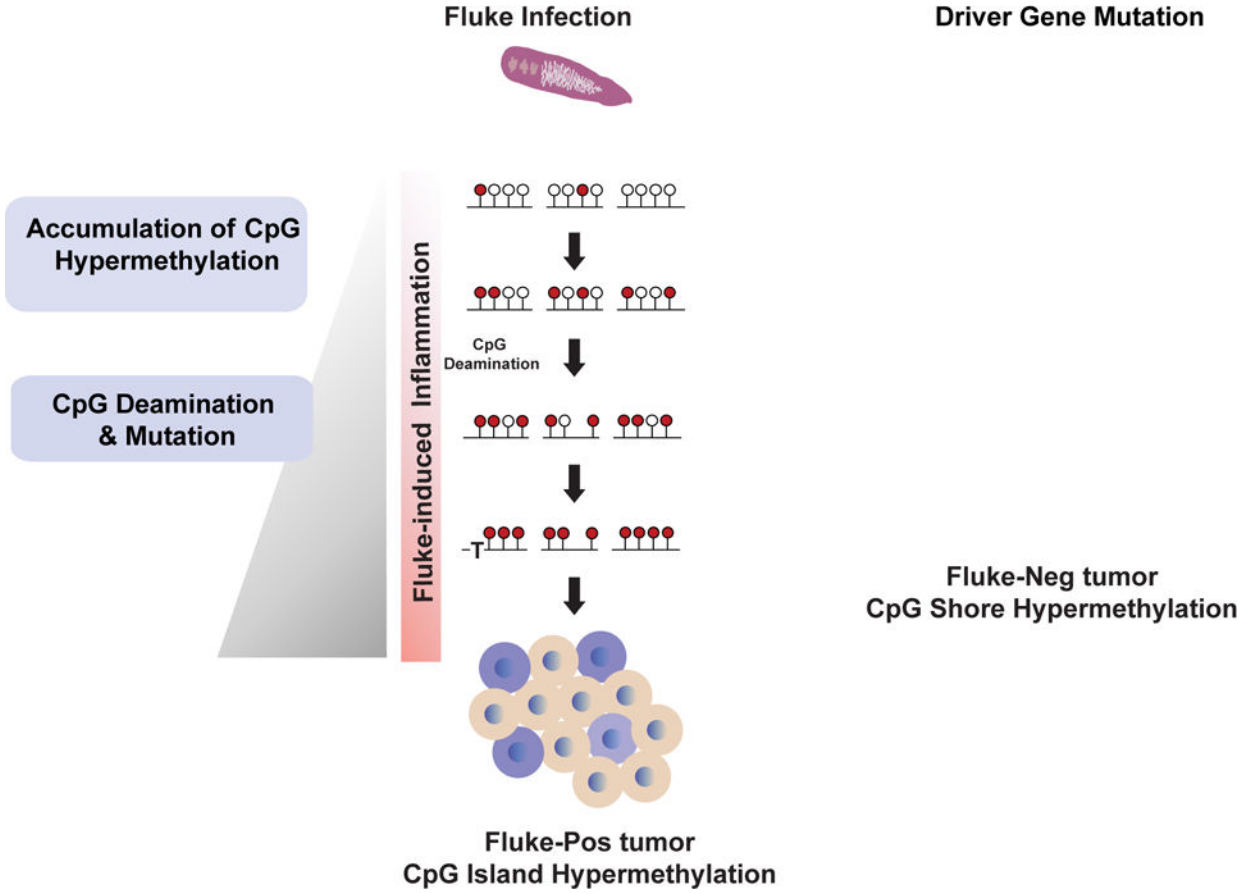

Figure 6. Model for Distinct Pathways of CCA Tumorigenesis

A proposed model for CCA development in Clusters 1 and 4 being driven by distinct mechanisms. Cluster 1 may be initiated by extrinsic carcinogens (fluke-infection) causing genome-wide epigenetic derangement and subsequent spontaneous 5-methylcytosine deamination and CpG>TpG mutations. In contrast, in Cluster 4 CCAs, intrinsic genetic mutations in strong driver genes such as *IDH1* reflect a primary initiating event and consequently drive DNA hypermethylation. See Discussion for details.