Supplementary Methods

Somatic mutation detection

We aligned sequence data to the human reference genome (hs37d5) using BWA-MEM v0.7.9a (1). We removed PCR duplicates using SAMTools (2) and performed indel realignments using Genome Analysis Toolkit v1.0 (3) (GATK). We used realigned data as input to both GATK Unified Genotyper and MuTect (4) to call sSNVs. We removed common polymorphisms based on dbSNP v132 (http://www.ncbi.nlm.nih.gov/SNP) and the 1000 genomes databases (http://www.1000genomes.org/), but retained any variants present in COSMIC v52 (http://www.sanger.ac.uk/cosmic). We applied filters to retain only high confidence sSNVs: variant depth ≥5, depth in matched normal sample ≥10, quality by depth ≥3, not present in a segmental duplication or repeat region (based on the UCSC annotations (5)). We also checked the depth of each variant allele in the matched normal sample to effectively remove any possible germline variants: in cases where the variant allele was present in more than 2 reads, we considered the variant to be germline and excluded it. Remaining variants were annotated using several gene annotation databases (CCDS, RefSeq and UCSC).

For indel identification, we used GATK IndelGenotyperV2 separately on tumors and matched normal samples, performing an electronic subtraction to retain candidate somatic variants. The set of indel filters included: variant depth ≥5 and ≥1 on each strand, variant present in at least 5% of the reads, depth in the normal sample ≥10, consensus average mapping quality of reads with the indel ≥55, the average base quality of bases in and around the indel ≥25, the average number of mismatches in reads with the indel ≤4, the average fraction of mismatched bases in reads with the indel ≤0.2, not present in a segmental duplication or repeat region, and not present in dbSNP. As indels usually have a very high false positive rate, we also manually inspected exonic and splice-site indels present in recurrently mutated genes using the Integrative Genomics Viewer (IGV) (6), retaining only variants passing visual inspection. We used the same methods for the targeted validation set (388 tumors), adding an additional filter that requires the variants to be inside the selected targeted regions. Confirming our analysis pipeline, orthogonal validation resequencing on 97 randomly selected sSNVs and 85 indels using Ion Torrent or Sanger technologies determined accuracy rates to be 99% for sSNVs and 87% for indels.

To check if different sequencing platforms (WGS, WES, or targeted sequencing) might bias our mutation calls, we performed three checks. (i) We directly compared ten samples with both WGS and WES data to one another, within the consensus targeted regions across the

sequencing platforms used. Reassuringly, for these ten samples, the average Jaccard's coefficient for mutation detection was 0.873, indicating a high degree of similarity in the specific mutations called. (ii) To assess if sequencing technique might affect overall mutation frequencies, we modelled the mutation frequency of each sample (dependent variable) as a function of sequencing technique, fluke status, and anatomical subtype as independent variables, using multivariate regression. We performed this test across 278 of the 404 targeted genes that were mutated in at least 1% of samples. We found that sequencing technique was not associated with overall mutation frequency ($p > 0.17$). (iii) To further check if mutations in individual genes might be biased towards specific sequencing techniques, we modelled for each targeted gene its mutation call in each sample (dependent variable) as a variable of sequencing technique, fluke status, and anatomical subtype (independent variables), using multivariate regression. We performed this test across 278 of the 404 targeted genes that were mutated in at least 1% of samples. We found that sequencing technique was also not associated with whether a mutation is called, in every gene checked (all $q > 0.19$, with multiple hypothesis correction).

Copy-number analysis

175 tumor-normal pairs were hybridized to Illumina HumanOmniExpress BeadChip arrays (SNParray) according to the manufacturer's instructions. The raw data was processed using Illumina Genome Studio to compute the $\log_2$R ratio (LRR) and the B-allele frequency (BAF). We used ASCAT v2.0 (7) to estimate tumor content and allele-specific copy-number profiles from SNP arrays. Using the generated profiles, we determined regions of copy-number alteration based on their relative copy-number (defined as the total copy number of the region divided by the average ploidy of the tumor) using the "copynumber" (8) Bioconductor package. Regions of copy-number loss were defined as regions with relative copy-number <0.7, while regions of copy-number gain were defined as the regions with relative copy-number >1.5. Regions of homozygous deletion were defined as regions with total copy-number equal to 0. Regions subject to loss of heterozygosity were regions in which one allele had copy-number 0.

For WGS tumor-normal pairs without SNP array data we estimated allele-specific copy-number profiles based on sequencing data using Control-FREEC (9). For the exome and targeted pairs, we estimated LRRs using Quandico (10). As *ERBB2* amplification is of therapeutic relevance, copy number analysis on *ERBB2* was interrogated further with one more method, Sequenza (11), on the WGS data.

We used GISTIC v2.0.22 (12) to determine regions of significant focal copy-number alterations in the WGS cohort ($q < 0.1$). Input segmentation files were generated based on ASCAT/Sequenza's inferred copy-number segments, and associated copy-number values were defined as $\log_2$ of the segment's relative copy-number.

To derive amplified genes in Cluster 3 (of our integrative clustering), we used the GISTIC focal_data_by_genes.txt output, and selected genes amplified (LRR ≥ 0.5) in more than half of Cluster 3 samples, and that showed statistical enrichment of amplified samples in Cluster 3 (Fisher's exact test, $q < 0.1$). Amplified chromosome arms in Cluster 3 were derived analogously. We chose not to use GISTIC to directly determine regions of significant amplifications in Cluster 3, due to its small sample size (7 samples).

Gene expression analysis

118 tumors were hybridized to Illumina HumanHT-12 Expression BeadChip arrays according to the manufacturer's instructions. Data was pre-processed and normalized using the "lumi" (13) R package with default parameters (quantile normalization). Probes that were not expressed or lacked gene annotations were removed, as were probes annotated as "Bad" or "No match" in the R annotation package "illuminaHumanv4.db". The latter probes have been determined to be unreliable (14) and correspond to low expressed probes or probes that are artefactually highly expressed due to non-specific hybridization. Batch effects were removed using ComBat (15). To analyse the relationship between somatic copy number at transcription level, expression values were first averaged across all probes in the respective gene. Benjamini-Hochberg method was used to adjust the p values from Wilcoxon rank-sum test for all tests done on expression data.

We used GSEA v2.2.2 (16) 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 (16).

Immune cell infiltration analysis

ESTIMATE (17) was used to determine the presence of infiltrating immune cells, using the ImmuneSignature geneset, in 118 tumors with gene expression data. A total of 126 genes were used to determine the immune score for each tumor.

DNA methylation analysis

DNA methylation profiles were obtained using the Infinium HumanMethylation450 BeadChip (450k array) for 138 tumors and 4 normal samples. Samples were hybridized to the arrays according to the manufacturer's instructions. We performed Noob background correction and BMIQ normalization to preprocess the raw data, using the "minfi" (18) and "wateRmelon" R packages (19). We removed probes with high detection p-values ($p > 0.05$) in any sample, probes corresponding to SNPs, and probes on sex chromosomes, resulting in 452,034 remaining probes. β-values and M-values were used for methylation analysis. The β-value is the ratio of the methylated signal versus the sum of methylated and unmethylated signal. The M-value is the log-ratio of the methylated versus unmethylated signal. We selected 1% (4,520) of the remaining probes on the chip for clustering. Among probes with mean $\beta < 0.5$ in the normal samples we selected the 4,520 with the highest standard deviations in β-values across the tumors. We clustered the tumors using the "RPMM" package (20) and visualized them with the "heatmap.2" function. To assess the robustness of the clusters, we repeated clustering using the top 0.5% and 2% of probes. Results were similar, with > 90% overlap in cluster membership.

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. Hypomethylated CpGs were defined analogously. We considered a genomic feature (such as a promoter region, a CpG island, or a CpG shore) to be hyper- or hypomethylated if at least one CpG site in that feature was hyper- or hypomethylated. Annotations of CpG probes to genomic features were obtained from the "IlluminaHumanMethylation450kmanifest" package (21). CpGs annotated with TSS1500, TSS200, 5'UTR, or First Exon relative to a gene (corresponding to the region 1500bp upstream of its TSS up to and including its first exon) were considered to occupy the gene's promoter region.

We also investigated the impact of DNA methylation on gene expression in 110 CCAs with both DNA methylation and gene expression data. Within each hypermethylated cluster, a gene was considered downregulated by promoter hypermethylation if: (1) at least one associated CpG in its promoter region was hypermethylated and (2) the CpG β-value exhibited

significant negative correlation with at least one expression probe ($q < 0.05$, Spearman's rank correlation). Hypomethylation and upregulation were analyzed analogously.

To explore associations between methylation clusters and genetic alterations or gene expression, we used either multivariate logistic regression with Firth bias correction from the "logistf" R package (22) (for categorical dependent variables) or multivariate rank regression from the "Rfit" R package (23) (for numerical dependent variables), and included fluke association and tumor subtype as variables to adjust for confounding. All p-values derived from multiple hypothesis testing were adjusted using the Benjamini-Hochberg method.

GSEA was used to identify enriched pathways among hypermethylated genes in each cluster. For each analysis, the input data was a list of genes, scored by the difference in group mean β-values of their most differentially-methylated promoter-region CpG sites. GSEA was run in pre-ranked mode, with the classic weighting scheme, on the MSigDB C2 catalogue of chemical and genetic perturbations.

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. We used Fisher's Exact Test to test for associations between mutations and hypermethylated CpGs.

Integrative clustering and clustering using individual platforms

iClusterPlus (24) 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 (7), (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). To evaluate the robustness of the clustering, we further ran 10 rounds of randomized subsampling clustering (taking 90% of features and 90% of tumors randomly and re-running iClusterPlus.tune in each round), and

confirmed that the optimal number of clusters chosen was also n=4, and the cluster assignments were also concordant with our clustering (>95% overlap in cluster membership).

We also performed unsupervised clustering on each of the four data platforms individually. Gene expression data was clustered by hierarchical clustering with the Ward.D2 method and Pearson correlation for distance. sCNA data was clustered by hierarchical clustering with the Ward.D2 method and Manhattan distance. Mutation data was clustered by hierarchical clustering with Ward.D2 method and Manhattan distance. Clustering for methylation data is described in the DNA methylation analysis subsection.

Data reanalysis based on expanded integrative clustering

We derived an expanded set of clustered samples by running iClusterPlus on the original 94 samples, together with samples missing one or more of the four data platforms. We ran iClusterPlus 7 times, each time on the original 94 samples together with (1) samples missing mutation data; (2) samples missing sCNA data; (3) samples missing expression data; (4) samples missing mutation and sCNA data; (5) samples missing mutation and expression data; (6) samples missing sCNA and expression data; (7) samples missing mutation, expression, and sCNA data. Methylation data was included in every run, as clustering without methylation data was unable to retrieve the original clusters. In each run, we used the original 94 samples to gauge the accuracy of the clustering, by evaluating their new cluster assignments against their original cluster assignments (which were considered the correct classification). Specifically, for each cluster in a run, its precision was estimated as the number of original samples *correctly* assigned to that cluster / the number of samples assigned to that cluster (ie. true positives / positives, where correct means matching the original cluster assignment). If a sample with missing data was assigned to a cluster with precision >0.9 in some run, we then assigned the sample to that cluster in the expanded clustering. The final expanded clusters consisted of 121 samples, made up of Cluster 1 (n=31: 27 original + 4 new), Cluster 2 (n=39: 25 original + 14 new), Cluster 3 (n=7: 7 original), and Cluster 4 (n=44: 35 original + 9 new) (Supplementary Fig. 1B).

Using the expanded integrative clusters, we performed the following data reanalysis:

i. In the analyses of associations between clusters and anatomical subtype, the expanded clusters confirmed that Clusters 1 and 2 were enriched in extrahepatic tumors while Clusters 3 and 4 were enriched in intrahepatic tumors ($p = 2.4x10^{-6}$ vs $1.2x10^{-5}$ in the original

clusters, Fisher's exact test). This persisted after adjusting for fluke status ($p = 2.1x10^{-5}$ vs $3.6x10^{-5}$ in the original clusters, multivariate regression).

ii. In the analyses of associations between clusters and fluke association, the expanded clusters confirmed that Cluster 1 was enriched in Fluke-Pos tumors ($p = 4.7x10^{-13}$ vs $1.1x10^{-11}$ in the original clusters, Fisher's exact test), and Clusters 3 and 4 were enriched in Fluke-Neg tumors ($p = 1x10^{-8}$ vs $3.8x10^{-9}$ in the original clusters, Fisher's exact test).

iii. In the analysis of association between clusters and mutations, the expanded clusters confirmed that: Clusters 1 and 2 were enriched in *TP53* mutations ($p = 2.5x10^{-8}$ vs $1.3x10^{-7}$ in the original clusters, Fisher's exact test); Cluster 1 was enriched in *ARID1A* mutations ($p = 0.0013$ vs 0.0071 in the original clusters, Fisher's exact test); Clusters 1 and 2 were enriched in *BRCA1/2* mutations ($p = 0.036$ vs 0.019 in the original clusters, Fisher's exact test).

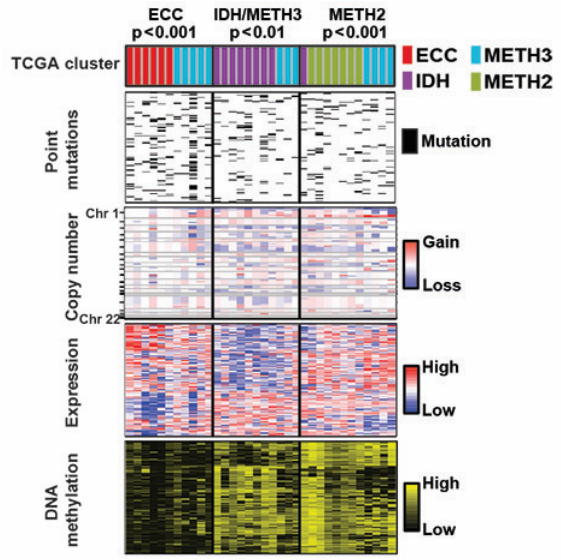
Samples newly-classified in the expanded clusters were also used to assemble a validation cohort for survival analysis (see Survival Analysis subsection).

Analysis of TCGA CCA data

We downloaded CCA data published from the TCGA group (25) from https://gdac.broadinstitute.org/ (26) and http://www.cbioportal.org/ (27), including somatic SNV data, sCNA data, RNAseq data, and DNA methylation data (TCGA level 3-processed). To confirm concordance in our analysis methods, we first ran these TCGA samples through our integrative clustering pipeline. To prepare the data for our pipeline, DNA methylation data, somatic mutations data, and sCNA data were processed (as needed) uniformly with our own samples. RNAseq data was processed by removing genes expressed in <20% of samples and liver-specific genes, using a list published by (25). We replaced expression values greater than 2 standard deviations from the mean with a value of 2 standard deviations from the mean, and applied a log2 transform. We retained genes with standard deviation >2 for clustering (1600 genes). We then performed integrative clustering by iCluster as we did for our samples. Our analysis identified three clusters largely overlapping the TCGA authors' own clusters (Supplementary Methods Fig. 1). The first cluster was associated with the authors' "ECC" cluster (extrahepatic CCA; $p < 0.001$, Fisher's exact test), being enriched in extrahepatic CCAs ($p < 0.1$). The second cluster was associated with the union of the authors' "IDH" and "METH3" clusters (*IDH*, *BAP1*, and *FGFR2* mutants; $p < 0.01$), exhibiting enrichment in samples with *IDH*, *BAP1*, and *FGFR2* mutants ($p < 0.01$). The third cluster was associated with the authors' "METH2" cluster (*CCND1*-amplified, $p < 0.001$), enriched in *CCND1*-amplified

samples ($p < 0.05$). Even though our pipeline did not split the *IDH* and *BAP1/FGFR* samples as the TCGA authors' pipeline did, our clusters nonetheless were not conflicting. This result lends reassurance to the robustness of our respective clustering pipelines.

We also related the TCGA clusters to the clusters derived from our cohort, based on the salient features describing each cluster. We found that the TCGA "IDH" (IDH mutants) and "METH3" (*BAP1* mutants and *FGFR* rearrangements) clusters matched our Cluster 4, which was also characterized by *IDH* and *BAP1* mutants, and *FGFR* rearrangements. The TCGA "ECC" cluster also matched our Cluster 2, which contained our fluke-negative extrahepatic CCAs. On the other hand, the TCGA "METH2" cluster (characterized by *CCND1* amplifications) did not match our clusters; while conversely our remaining fluke-negative group Cluster 3 also did not map to any of the TCGA clusters. Finally, since our Cluster 1 was a fluke-positive group, it naturally did not match any of the TCGA clusters (since the TCGA samples are exclusively fluke-negative). Taken collectively, these results suggest that most of the TCGA clusters, especially their main cluster of study ("IDH"), are concordant with our ICGC clusters, and that neither classification strictly precludes the other. The TCGA CCA samples were also used to build a validation cohort for survival analysis (see Survival Analysis subsection).



Supplementary Methods Figure 1. Clustering of TCGA CCA samples via our pipeline identified three clusters overlapping the original TCGA clusters.

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 for the integrative CCA clusters, we also analysed a separate validation cohort of 58 samples, by combining two sources:

i. 25 samples (with survival data) that were newly classified into CCA clusters under the expanded integrative clustering approach (see above Data reanalysis based on expanded integrative clustering subsection for details);
ii. 33 recently-published CCA samples with survival data from Farshidfar et al. (2017) (25). Three subgroups proposed by the authors closely matched our own CCA clusters in terms of salient features: their "IDH" and "METH3" subgroups, distinguished by *IDH* and *BAP1* mutants and *FGFR* rearrangements, matched our Cluster 4; and their "ECC" subgroup, composed of extrahepatic fluke-negative CCAs, matched our Cluster 2. To validate the robustness of these subgroups, we also subjected the Farshidfar et al. (2017) (25) samples through our own integrative clustering pipeline, and obtained similar clusters (see above Analysis of TCGA CCA Data subsection for details). We thus added these samples as Cluster 2 (n=5) and Cluster 4 (n=23) in our validation cohort (the remaining 5 samples belonged to an unmatched subgroup "METH2").

Besides the CCA clusters, we also tested for survival associations against various cancer stem cell markers (*CD133*, *CD13*, *CD44*, *SOX2*, *CD90* and *Nanog*). Of these, only *CD13* significantly correlated with poorer prognosis in our cohort ($p < 0.05$, log-rank test). However, this was not significant in a multivariate survival analysis after accounting for fluke status, tumor staging, and tumor grade.

Driver gene analysis

We integrated (i) 71 CCAs WGS (average depth 66x, Supplementary Table 1C), (ii) a targeted sequencing cohort of 188 CCAs surveying 404 genes (>500x average coverage; Supplementary Tables 5B-C) reported as mutated in various hepatobiliary cancers (28-34), and (iii) a recently published Japanese exome cohort of 200 cases (34) (>100x average coverage; Supplementary Table 5B) and performed gene significance analyses using MutSigCV (35) and IntOGen (36). 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.

Detection and annotation of structural variations

BWA-MEM alignments from each tumor-normal pair were analyzed by CREST (Clipping REveals STructure) (37). Preliminary SVs deemed as ambiguous by PTRfinder (38) were discarded. 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.

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) (39). Only SVs with ≥ 2 reads with poly-A tails at the putative L1 insertion site were retained for further analysis. Circos plots were generated using "Circos" (40).

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 (Supplementary Table 4) by PCRs using primers flanking predicted sites of insertion. PCRs were performed using AccuPrime™ Pfx DNA Polymerase (Invitrogen) with 200 ng of WGA DNA.

Mutation signature analysis

Trinucleotide-context mutation spectra were generated by counting the number of the 96 somatic substitution types (C:G>A:T, C:G>G:C, C:G>T:A, T:A>A:T, T:A>C:G and T:A>G:C in each trinucleotide context, i.e. the context of the immediately flanking nucleotides) for each

tumor. NMF was applied to the trinucleotide-context mutation spectra of CCAs using published software (41) to extract mutation signatures. Six stable and reproducible mutational signatures were extracted from CCAs and termed signatures A, B, C, D, E, F. These 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. Signatures A, B, C, E, F resembled Signatures 1, 2, 22, 17, and 20, respectively, with cosine similarities > 0.9. Signature D resembled Signature 5 with a cosine similarity of 0.85. To evaluate the contributions of the COSMIC signatures to mutations in CCA, we used supervised NMF by introducing these 6 COSMIC signatures plus Signature 13, which always co-occurs with Signature 2. Visual inspection showed that 5 CCAs had strong transcription strand bias in T>C mutations and enrichment for T>C mutations in ATN contexts, both features of Signature 16. Three CCAs had high ratios of CC>AA mutations to C>A mutations, features of Signatures 4 and 8. Visual inspection also revealed a Signature 6-like mutation pattern (enrichment for CCT>CAT mutations) in 2 of the 3 MSI cases and an aflatoxin-like mutation pattern (Signature 24) in one sample. We re-analysed these samples individually with the additional possible signatures and chose signature combinations that substantially reduced the Euclidean distances between the observed and reconstructed spectra. 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. To test enrichments of mutation signatures in CCA subtypes, multivariate rank regression was applied using R function “rfit” from the “Rfit” package, where fluke association, anatomical subtype, and patient age were included as variables to adjust for confounding. The MSI status in the prevalence set was determined by the indel counts in simple repeat sequences. MSI samples showed higher counts (≥ 6 indels) while other samples have low counts (≤ 3 indels).

Analysis of somatic promoter mutations with FIREFLY

FIREFLY (FInding Regulatory mutations in gEne sets with FunctionaL dYsregulation) is a novel method that identifies gene sets dysregulated by somatic promoter mutations through modulation of transcription factor (TF) binding. Compared to approaches used in previous cancer genome studies (42,43) FIREFLY differs in three important respects. First, unlike other studies using positional weigh matrices (PWMs) (44) FIREFLY uses experimentally determined high-throughput TF-DNA binding data (45,46) generated by protein-binding microarray (PBM) assays to predict mutation-associated changes in TF binding affinity. This approach overcomes shortcomings of traditional PWM models, which cannot capture multiple

modes of binding or interdependencies within TF binding sites. Second, FIRELY 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, FIREFLY then orthogonally validates the functional impact of the binding-change predictions using expression data of primary tumors.

We extracted the set of somatic promoter mutations from 70 WGS tumors (after excluding 1 MSI tumor) by selecting non-coding sSNVs within +/- 2kb of TSSs of GENCODE genes. We identified those mutations predicted to change TF-binding based on universal protein-binding microarray (PBM) data (details below).

We evaluated 1,150 gene sets for enrichment in promoter binding-change mutations. These gene sets included the KEGG and REACTOME gene sets in MSigDB as well as gene sets of interest in CCA, based on our DNA methylation analysis. 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 a gene sets passing these tests, we then performed gene expression tests to assess their transcriptional dysregulation.

Fisher's exact test was applied to assess the significance of M, based on the contingency table of genes with and without binding-change mutations, within and outside of the gene set (Fig. 3B, left panels).

For the synthetic mutations test, we created 1000 sets of 70 tumors with synthetically-generated mutations. We generated synthetic mutations in promoter regions based on the genome-wide per-trinucleotide mutation frequencies observed for each tumor/normal pair (details below) (43,47). For each set of synthetic tumors, we calculated the test statistic, M', analogous to M for the set of real tumors. We took the distribution of M' over the 1,000 sets of synthetically mutated tumors as the null distribution for M (Fig. 3B, center panels).

For the gene expression test, we used the "GSA" v1.03 R package (48). For each of the gene sets enriched in binding-change mutations, we asked whether there is an association between the number of binding-change mutations in that gene set, and its genes' expression dysregulation. We ran GSA for each gene set separately, in quantitative mode without restandardization, and adjusted the P-values using the Benjamin-Hochberg method. To obtain gene expression boxplots (Fig. 3B, right panels), for each upregulated gene set we selected the

genes with GSA scores > 1.5 (< -1.5 for downregulated), and plotted their z-normalized gene expression values.

We additionally checked if the number of dysregulated gene sets among those enriched in binding-change mutations was higher than among randomly-selected gene sets. We generated 100 sets of randomly-selected gene sets, with each set following the number and sizes of the gene sets enriched in binding-change mutations. We ran GSA on these randomly-selected sets, and counted the number of dysregulated gene sets in each set (Supplementary Fig. 6C).

Identifying TF binding-change mutations

To identify binding-change mutations, we used TF-DNA binding specificity data from universal protein-binding microarray (PBM) experiments, downloaded from the cis-BP database (45) 998 PBM data sets for 486 mammalian TFs, covering a broad range of TF families). Universal PBM experiments measure TF binding to all possible 8-mer sequences with a DNA binding enrichment score (E-score) (46). E-scores are derived from a modified form of the Wilcoxon-Mann-Whitney statistic, and range from -0.5 to +0.5, with higher values corresponding to higher TF binding affinity. Typically, E-scores > 0.35 correspond to specific TF-DNA binding (49).

To call binding sites for a particular TF, we used a stringent E-score cutoff of 0.4, which corresponds to a false discovery rate of < 0.001 (50). To further increase our confidence in the identified TF binding sites, 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, i.e. regions for which we have high confidence that there is no potential for specific TF-DNA binding.

For each somatic mutation and each TF with available PBM data, the method analyzes 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 is 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 is 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 486 TFs with available PBM data in the cis-BP database.

Generating synthetic mutations based on trinucleotide mutation frequencies

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. We first computed the frequency of each triplet across the normal genome, on either strand, e.g.:

P(*w*=ACG) = #_of_occurrences_of_ACG_in_reference_genome / genome_size.

For each tumor pair, we then computed the mutation density across the genome:

P(*w*≠*t*) = total_#_of_mutations / genome_size.

Next, we computed the proportion of each particular mutation type (e.g. ACG>AAG) among all mutations:

P(*w*=ACG,*t*=AAG | *w*≠*t*) = #_of_ACG_to_AAG_mutations / total_#_of_mutations,

where *w* and *t* are the triplet sequences in the normal and tumor, respectively. The probabilities were computed from the entire genome because promoters harbour too few mutations to allow accurate probability estimates based on promoter sequences alone. Then the probability of seeing a particular triplet sequence (e.g. ACG) at a particular mutation is

P(*w*=ACG | *w*≠*t*)

= P(*w*=ACG,*t*=AAG | *w*≠*t*) + P(*w*=ACG, *t*=AGG | *w*≠*t*) + P(*w*=ACG, *t*=ATG | *w*≠*t*).

Using Bayes' rule, we can write the probability of generating a mutation at a triplet (e.g. ACG) as

P(*w*≠*t* | *w*=ACG) = P(*w*=ACG | *w*≠*t*) * P(*w*≠*t*) / P(*w*=ACG).

To generate synthetic mutations, we iterated through all positions in the promoter regions of interest and at each position we did the following:

1. Determine whether the nucleotide should be mutated, given the identity of the triplet centered at that position, according to the conditional probabilities computed above (e.g. **P(*w*≠*t* | *w*=ACG)**).
2. If the nucleotide was selected to be mutated, then it was mutated according to the corresponding frequencies computed from the tumor-specific mutations. E.g.:

   P(*t*=AAG | *w* =ACG, *w*≠*t*)

   P(*t*=AGG | *w* =ACG, *w*≠*t*)

   P(*t*=ATG | *w* =ACG, *w*≠*t*)

This procedure can be followed to generate tumor-specific synthetic mutations in any set of genomic regions of interest. Here, we focus on generating mutations in promoter regions.

Given that the marginal probability of a position being mutated, i.e. **P(*w*≠*t*)**, may be different for promoter regions compared to genome-wide, in the procedure described above we defined

P(*w*≠*t*) = total_#_of_mutations_in_promoters / total_size_of_promoters

rather than

total_#_of_mutations / genome_size.

SUPPLEMENTARY FIGURE LEGENDS

Supplementary Figure 1. Unsupervised Clustering on CCA samples.

(A) Robustness of CCA Clusters from randomized subsampling clustering.

(B) Expanded clusters (121 samples) by including samples with one or more missing data platforms, while retaining cluster prediction accuracy of 90%.

(C) Clustering on gene expression, copy number alteration, and point mutation. Clustering on DNA methylation is shown in Fig. 4A.

(D) Clustering on tumors stratified by anatomical location.

Supplementary Figure 2. Alterations Found in CCA Clusters.

(A) Mutation burdens in Clusters. "POLE" indicates a case with DNA polymerase epsilon proofreading deficiency. MSI status was defined by indel counts (≥ 6 indels) in simple repeat sequences.

(B) and (C) Gene expression of CTNNB1, WNT5B and AKT1 (B) and ERBB2 (C).

(D) Immune score representing the infiltration of immune cells in tumor tissues.

(E) Heatmap of immune-related gene expression in Clusters. P-values were computed using the Wilcoxon rank-sum test.

(F) Survival graphs for independent validation cohort, Fluke-Pos vs. Fluke-Neg CCAs, and intrahepatic vs. perihilar vs. distal CCAs.

Supplementary Figure 3. MAP2K4 Homozygous Deletions and ERBB2 Amplifications.

(A) and (B) Regions of homozygous deletion in two CCA tumors. Left side shows a region at Chromosome 17p containing the MAP2K4 gene. Right side shows another chromosomal region for comparison purposes.

(C) Significant decreases in expression level associated with samples where MAP2K4 is homozygously deleted. P-values were computed using one-sided Wilcoxon rank-sum tests.

(D) Validation of ERBB2 amplifications in two CCAs by FISH. FISH was performed using ERBB2 (red)/CEP17 (green) probe sets and performed on paraffin-embedded tissue.

Supplementary Figure 4. Alterations Related to Structural Variations Found in CCAs.

(A) 71 CCA whole-genomes and association of SV burden with genomic alterations/Fluke status.

(B) CIRCOS plot of predicted somatic L1-retrotransposition events.

Supplementary Figure 5. FIREFLY Analysis of Pathways Systematically Dysregulated by Somatic Promoter Mutations that Alter Transcription Factor Binding.

(A) Details of two example non-significant gene sets.

(B) Null distribution of the number of dysregulated gene sets found by GSA out of 19 randomly-selected gene sets.

(C) Relative luciferase activity (mutant/wildtype) for binding-change mutations in promoters of PARD3, PIAS1, AICDA are shown in two cell lines (H69 and EGI1). Values represent means ± S.D. in 3 biological replicates. P-values were derived from one-sample t-test (testing for mean ≠ 1.0).

(D) Differential distribution of binding-change mutations in three enriched gene sets across the four CCA clusters.

Supplementary Figure 6. Epigenetic Clusters and Mutation Signatures.

(A) Starburst plots of gene promoter CpG methylation and gene expression in Cluster 1 (left) and Cluster 4 (right). Points represent genes with significant negative correlations between promoter CpG methylation and gene expression (red), genes with significant positive correlation (blue), and genes with no correlation (grey). Only genes with significantly hypermethylated or hypomethylated promoter CpGs are shown. Accompanying pie charts show the percentages of hypermethylated gene promoters with significant correlations between promoter CpG methylation and gene expression.

(B) TET1 and EZH2 expression levels.

(C) Volcano plot showing association of BAP1 alterations with CpG hypermethylation.

(D) Contributions of mutation signatures to CCA clusters.

(E) Contingency tables show associations between mutations and methylation status in Cluster 1 (left) and Cluster 4 (right), for CpG > TpG mutations (top) and non C > T mutations (bottom). Only mutations near a CpG probe (within 50 bp), which is unmethylated in normal, were considered.

SUPPLEMENTARY TABLE LEGENDS

Supplementary Table 1. Summary of Patients and Clinical Data.

Supplementary Table 1A. Detailed patients data, including clinical information, available genomic data, and CCA cluster membership.

Supplementary Table 1B. Summary statistics of data.

Supplementary Table 1C. Coverage statistics for the 71 tumor-normal pairs analysed by whole-genome sequencing.

Supplementary Table 2. Alterations in CCA Clusters.

Supplementary Table 2A. Aberrant pathways in each cluster based on GSEA analysis.

Supplementary Table 2B. Amplified chromosomal regions and genes in Cluster 3.

Supplementary Table 2C. Univariate and multivariate Cox proportional hazards analysis of overall survival for tumors in Clusters.

Supplementary Table 3. Summary of CCA Alterations.

Supplementary Table 3A. List of somatic nonsilent single nucleotide variations and indels detected across the 404 targeted genes.

Supplementary Table 3B. List of genes affected by recurrent nonsilent somatic mutations across all 459 tumors.

Supplementary Table 3C. Significantly mutated genes identified by both MutSigCV and Intogen.

Supplementary Table 3D. Summary table of mutations and their prevalence.

Supplementary Table 3E. Samples with ERBB2 amplification.

Supplementary Table 3F. Activated pathways among samples with ERRB2 amplification.

Supplementary Table 3G. Other amplifications and deletions (MYC, MDM2, EGFR, CCND13 amplifications and CDKN2A, UTY, KDM5D deletions).

Supplementary Table 4. Structural Variations across 71 WGS CCAs.

Supplementary Table 5. Coverage Statistics and List of Genes for Targeted Sequencing.

Supplementary Table 5A. Number of somatic nonsilent mutations per tumor in 71 WGS.

Supplementary Table 5B. Coverage statistics for 388 tumor-normal pairs.

Supplementary Table 5C. Gene list used for targeted sequencing.

Supplementary References

1. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010;26:589-95.
2. 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.
3. 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.
4. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 2013;31:213-9.
5. Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, Diekhans M, *et al.* The UCSC Genome Browser database: 2015 update. Nucleic Acids Res 2015;43:D670-81.
6. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, *et al.* Integrative genomics viewer. Nat Biotechnol 2011;29:24-6.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.

13. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. Bioinformatics 2008;24:1547-8.
14. Barbosa-Morais NL, Dunning MJ, Samarajiwa SA, Darot JF, Ritchie ME, Lynch AG, *et al.* A re-annotation pipeline for Illumina BeadArrays: improving the interpretation of gene expression data. Nucleic Acids Res 2010;38:e17.
15. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 2007;8:118-27.
16. 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. Proc Natl Acad Sci U S A 2005;102:15545-50.
17. 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.
18. 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.
19. 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.
20. 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.
21. Hansen KD, Aryee MJ. IlluminaHumanMethylation450kmanifest: Annotation for Illumina's 450k methylation arrays. R package version 040 2012.
22. Heinze G, Ploner M, Beyea J. Confidence intervals after multiple imputation: combining profile likelihood information from logistic regressions. Stat Med 2013;32:5062-76.
23. Kloke JD, McKean JW. Rfit: Rank-based Estimation for Linear Models. R Journal 2012;4/2.
24. 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.

25. 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.
26. BroadInstituteTCGAGenomeDataAnalysisCenter. 2016 Analysis-ready standardized TCGA data from Broad GDAC Firehose 2016_01_28 run. Broad Institute of MIT and Harvard <https://doi.org/10.7908/C11G0KM9>.
27. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Science signaling 2013;6:pl1.
28. 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.
29. 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.
30. 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.
31. 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.
32. Zou S, Li J, Zhou H, Frech C, Jiang X, Chu JS, *et al.* Mutational landscape of intrahepatic cholangiocarcinoma. Nat Commun 2014;5:5696.
33. 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.
34. 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.
35. 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.
36. 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.

37. 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 Meth 2011;8:652-4.
38. 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.
39. 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.
40. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, *et al.* Circos: an information aesthetic for comparative genomics. Genome Res 2009;19:1639-45.
41. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering signatures of mutational processes operative in human cancer. Cell reports 2013;3:246-59.
42. 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.
43. Melton C, Reuter JA, Spacek DV, Snyder M. Recurrent somatic mutations in regulatory regions of human cancer genomes. Nat Genet 2015;47:710-6.
44. Siggers T, Gordan R. Protein-DNA binding: complexities and multi-protein codes. Nucleic Acids Res 2014;42:2099-111.
45. 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.
46. 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.
47. Araya CL, Cenik C, Reuter JA, Kiss G, Pande VS, Snyder MP, *et al.* Identification of significantly mutated regions across cancer types highlights a rich landscape of functional molecular alterations. Nat Genet 2016;48:117-25.
48. Efron B, Tibshirani R. On testing the significance of sets of genes. 2007:107-29.
49. 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.

50. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, *et al.* Diversity and complexity in DNA recognition by transcription factors. Science 2009;324:1720-3.