Consensus molecular subtypes define distinct evolutionary trajectories of biliary tract cancers

Felix E.G. Beaudry[1,2], Duhan Yendi[2], Danielle Arshinoff [3,5], Nicholas Light[2], Simona Perrotti[2], Erin Winter[2], Liam R. Cristant[2], Amelia Xu[1], Julie Wilson[1], Anna Dodd[2], Roxana Bucur[2], Eric X. Chen[2, 3], Elena Elimova[2, 3], Rebecca Wong[2, 4], Aruz Mesci[, 4], Ali Hosni[, 4], Anand Ghanekar [2,5], Raymond Jang[2, 3], Chaya G Shwaartz[5,6], Trevor Reichman[5,6], Carol-Anne Moulton[5,6], Enrique Sanz Garcia[2, 3], Grainne M. O'Kane[7], Erica S. Tsang[2], Xin Wang[2], Ian McGilvray[5,6], Steven Gallinger[1,5,6], Trevor J. Pugh[1,2,8], Gonzalo Sapisochin[6,9], Arndt Vogel[2,6], Jennifer J. Knox[2], Faiyaz Notta[1,2,8], Robert C. Grant[1,2,3,10,*]

[1]Ontario Institute for Cancer Research, Toronto, Canada.

[2]Princess Margaret Cancer Centre, Toronto, Canada.

[3]Institute of Medical Sciences, University of Toronto, Toronto, Canada.

[4]Department of Radiation Oncology, University of Toronto, Toronto, Canada.

[5]Department of Surgery, University of Toronto, Toronto, Canada.

[6]Toronto General Hospital, University Health Network, Toronto, Canada.

[7]St.Vincent's University Hospital and School of Medicine University College Dublin, Ireland.

[8]Department of Medical Biophysics, University of Toronto, Canada.

[9]Vall d'Hebron University Hospital, Barcelona, Spain.

[10]ICES, Toronto, Canada.

*Correspondence: Robert.Grant@UHN.ca

Abstract

Biliary tract cancers are clinically and molecularly heterogeneous. To establish a unified molecularly-informed classification, we performed whole genome and transcriptome sequencing of 169 tumors enriched with laser capture microdissection and annotated with clinical records. Two consensus molecular subtypes (BTC-CMS) were revealed by network integration of existing transcriptomic classes, and could be reliably predicted from clinical genetic panels. BTC-CMS-B tumors exhibited *TP53*, *KRAS*, and *SMAD4* mutations, higher tumor mutational burden, APOBEC mutational signatures, and extrachromosomal DNA. These tumors spanned the biliary tree and were associated with chronic inflammation. BTC-CMS-A tumors carried *BAP1*, *IDH1*, and *FGFR2* alterations, occurred exclusively in the liver, and conferred superior survival after surgery and in advanced disease. Overall, the BTC-CMS framework defines two molecular lineages that outperforms patient stratification based on anatomy alone.

Introduction

Cancers with similar clinical presentations can differ profoundly at the molecular level, resulting in distinct clinical trajectories and therapeutic vulnerabilities. Molecular subtyping refines disease classification by identifying biologically coherent subgroups. Robustly defined subtypes serve as a roadmap for stratifying patient outcomes [1] and predicting sensitivity to precision [2,3] and systemic therapies [4,5].

Biliary tract cancer (BTC) groups a family of rare cancers with high recurrence rates after surgery and poor survival, especially in advanced disease [6,7]. BTCs are often considered together due to their anatomical proximity and shared clinical management, but this masks heterogeneity across risk factors, anatomic location, histopathology, genomics, and outcomes. The different staging across BTCs showcases its heterogeneity, with distinct systems for intrahepatic, perihilar, distal, and gallbladder cancers. Incorporating molecular characteristics could provide a unifying framework for the expanding subclassification of BTC.

Genomic classification of BTC in the clinic is currently based on individual gene mutations. While BTCs have high rates of genetic alterations targetable with precision therapies [8], each individual driver is rare [9,10], such that subtypes defined by individual actionable drivers are impractical and imprecise. Gene expression-based classifiers of BTC have found variable numbers of latent classes and the clinical associations have been inconsistent [11,12], potentially explaining their limited uptake. Because BTCs are highly desmoplastic tumors, classifiers based on expression inferences of the tumor immune microenvironment have been proposed for bulk specimens [13] but offer limited insight into the biology of the cancer cell itself. Overall, the rarity and high molecular heterogeneity of the cancer, and low tumor content of samples, has hindered efforts to understand BTC tumors at a molecular level.

Here, we unite findings from genomics [14], cell lines [15], single cell RNAseq [16] and clinical outcomes [17] to characterize two BTC consensus molecular subtypes (BTC-CMS; Fig 1a). We develop and test lightweight expression- and genetics-based classifiers that robustly distinguish these subtypes, allowing for novel integration of previously disparate public datasets. For deep dissection of genetic drivers, etiology, mutational signatures and expression profiles of the subtypes, we performed whole genome and transcriptome sequencing on 169 BTC tumor samples. This sample size is over two times larger than previous whole genome cohorts [11,12] and enriched with laser capture microdissection providing the first cohort to appropriately power *de novo* detection of genomic drivers in BTC [18,19]. Integration of our molecular and clinical data show BTC-CMS represent broad differences in the evolutionary trajectories of cancers and that tracking CMS will help clinical trial efficacy in BTC.

Results

Cohort description

We analyzed one tumor specimen for each of 169 patients with BTC, including 72 previously reported cases [8,20] and 97 new cases recruited prospectively. The median age of patients was 60 years (range 29-83) and tended towards males (n = 96). 97 patients underwent curative-intent resections, followed by adjuvant capecitabine (n

= 43), and 103 received palliative-intent systemic therapy, most commonly gemcitabine, cisplatin, and durvalumab (n = 65) (Table S1; Fig. 1b); clinical course information was unavailable for 12 patients.

Tumors spanned the biliary tree, including 104 intrahepatic, 11 mixed hepatocellular cholangiocarcinoma (HCC-CCA), 26 perihilar, 9 distal, and 18 gallbladder primary tumors (Fig. 1c). 27 specimens were taken from metastases, including liver biopsies from 19 extrahepatic cancers (Fig. 1d). Over half of the samples (n = 105, 62%) were treatment-naive (Fig. S1a). Because BTC typically exhibits desmoplasia and low tumor cellularity, most samples were enriched using laser-capture microdissection (n = 153; Fig. S1b). Fresh frozen samples (n = 148) were preferred over formalin-fixed paraffin-embedded tissue (n = 21). Tumor and normal samples underwent whole genome sequencing (WGS), achieving a median coverage of 85-fold for the tumor, and 38-fold for the normal. Tumors also underwent whole transcriptome sequencing (WTS), when enough tissue remained (n = 134), with a median of 84 million reads per sample.

Published transcriptomic classifiers converge on two consensus molecular subtypes

To explore overlap between previous expression-based classifiers (Table S2), we assigned expression classes to all 116 fresh frozen samples with WTS in our cohort (Fig. 2a, S2a). We constructed a similarity network in which nodes represented the gene expression classes and edge weights reflected how often samples were assigned to both of the classes connected by that edge, and performed Markov clustering on the network [21] (Fig. S2b). After network optimization (see *Methods*, Fig. S2c), ten classes from five classifiers collapsed into two significantly interconnected clusters (p = 0.002, hypergeometric test, Fig. 2b). We named these groups biliary tract cancer consensus molecular subtypes A (BTC-CMS-A) and B (BTC-CMS-B). Strikingly, signatures contributing to the CMS originated from distinct methods: Andersen *et al.* and O'Rourke *et al.* identified prognostic groups [17,22]; Vijay *et al.* performed unsupervised clustering of bulk WTS of BTC and hepatocellular cell lines [15]; and Song G. *et al.* developed a classifier from single-cell RNAseq of intrahepatic BTCs [16]. These findings highlight previously unappreciated agreement between existing expression classes, suggesting broad applicability.

To enable further exploration of subtypes across datasets and data types, we developed a parsimonious expression classifier (hereafter eCMS). We trained a top scoring pair model on 80% of the cohort using samples where all classifiers agreed on CMS cluster assignment (n = 70). We found 16 gene pairs could effectively distinguish samples from the two CMS (Fig. 2a). Testing the eCMS classifier in the remaining 20% of samples achieved 100% accuracy (95% C.I.: [82 - 100%]), with more than 12 of the 16 gene pairs agreeing on CMS assignment on all test samples (Fig. S2d).

We used protein expression data from Dong *et al.* [23] and methylation data from Jusakul *et al.* [11] to confirm the generalizability of the CMS. Classifying samples from Dong *et al.* [23] using the eCMS, the first principal component of variance in protein expression, explaining 10.8% of the variance among samples, was significantly higher in CMS-A than CMS-B (p = 2.87e-05, Fig. S2e). In methylation data from Jusakul *et al.* [11], the first principal component, explaining 10.8% of the variance, similarly separated by eCMS with values being significantly higher in CMS-B (p = 3.6e-12, Fig. S2f). These results showed the CMS were distinguishable across multiple molecular data beyond gene expression.

Overall survival differed significantly by CMS, linking prior observations across individual classifiers [17, 22]. In our cohort, CMS-A was associated with significantly longer overall survival with first-line palliative systemic therapy than CMS-B, adjusting for stage at presentation and anatomical location (p = 0.02; median survival of 22.4 months in CMS-A and 15.0 months in CMS-B; Fig 2c, d). In external datasets, CMS stratified survival after surgery in 147 iCCA patients from Dong *et al.* [23] accounting for disease stage (p = 9.5e-05) and survival from diagnosis for 62 BTC patients from Jusakul *et al.* [11] (p = 0.002) accounting for disease stage and anatomical location (Fig S2g, h). Together, these data indicate that prior classifiers of BTC converge on two robust CMS with prognostic importance.

Subtypes harbor distinct genomic drivers

Turning to our whole genome sequencing for all 169 samples, eleven genes were significantly enriched for small substitution mutations (< 2 codons in length) across both a coding-centric and a coding-agnostic approach (see *Methods*), with three additional genes only passing the false discovery rate (FDR) in one approach (Fig. 3a, Table S3). The two genes with the strongest signal of over-representation across both tests were *TP53* (n = 59, 35%) and *BAP1* (n = 30, 18%). Among gene promoters, we identified recurrent small substitutions in *TERT* (n = 7, FDR q = 0.01). Small substitutions driver analysis restricted within CMS-B recovered *CDKN2A* (n = 5, q = 0.01) and *ARID2* (n = 10, q = 0.01) as additional drivers and no additional drivers were found when the analysis was restricted to CMS-A cases (Fig. S3a).

Under Knudson's two-hit hypothesis [24], copy-loss events leading to loss-of-heterozygosity should converge on tumor suppressor genes. Six genes showed significantly more loss-of-heterozygosity events in the presence of a small substitution than when the gene was wild-type, with loss-of-heterozygosity occurring at the highest rates with substitutions in *TP53* (52 loss-of-heterozygosity events among 59 substitutions, FDR = 2.5e-15; Fig. S3b). All loss-of-heterozygosity events at *TP53* or *STK11* were copy-neutral. Loss-of-heterozygosity did not occur with mutated *IDH1* and *NRAS*, supporting their roles as oncogenes rather than tumor suppressor genes in BTC.

Next, we explored genes impacted by recurrent structural variants. To enrich for functional impact, we tested structural variant breakpoints as expression quantitative trait loci (eQTL) [25]. Top structural variant expression quantitative trait loci within CMS-A were *FGFR2* (n = 19), and *TERT* (n = 3), while top eQTL for CMS-B were *CDK6* (n = 8) and *MDM2* (n = 5) (Fig. S3c). Amplicon reconstructions from copy number profiles found the regions encompassing *MDM2* and *CDK6* were commonly extrachromosomal DNA (n = 4 and n = 3, respectively). Because most high effect size structural variant-eQTL were rare (<5% of cases), we formally assessed recurrence with a Q-Q enrichment analysis [26]. *FGFR2* was the only gene with somatic structural variants showing significant expression quantitative trait association and statistical recurrence (FDR = 2.4e-04). All *FGFR2* structural variants were confirmed to represent *FGFR2* fusions using RNA where available (n = 15). Three cases had small substitutions in *FGFR2* but had *FGFR2* expression levels similar to fusion cases (Fig. S3d), suggesting alternate mutations can yield the same phenotype.

Next, we explored how mutations in drivers were co-distributed among tumors (Fig. S3e). After correcting for mutation frequency, co-occurrences and mutual exclusivity patterns hinted at higher order associations (Fig. 3b). By assigning CMS associations to genes by testing how often samples from each subtype were mutated versus wildtype, we revealed a nexus of mutual exclusively defined by *TP53*, *STK11*, *SMAD4*, and *KRAS* mutations in CMS-B, in contrast to mutations in *NRAS*, *IDH1*, *FGFR2*, and *BAP1* in CMS-A (Fig. 3b, c). These results suggest driver mutations align with the CMS established using expression profiles.

Evaluating recurrence of copy number alterations [27], we found well-characterized oncogenes such as *ERBB2* and *MYC* as well as recurrent homozygous deletions of tumor suppressor gene *CDKN2A* (Fig. 3d). GISTIC within subtypes revealed divergent copy number profiles: recurrent narrow amplifications are common in CMS-B, including *GATA6*, *CCND1* and *CCNE1*, as well as *MDM2* and *CDK6*. In contrast, CMS-A commonly showed chromosome-arm scale heterozygous deletions (single arm losses). Tumor suppressor genes in other cancer types with changes in gene expression in our cohort in CMS-A include *ERFFI1* and *IRF2*, but we also note recurrent deletion of 6q and amplification of 1p21 without conclusively identifiable drivers.

Given the significantly different driver profiles of the subtypes, we next tested the capacity of a genetics-only classifier to predict CMS (hereafter genomic CMS or gCMS). To keep generalizability across platforms including many genetic panels, we trained a model on median arm-level relative copy number (*i.e.* log tumor / normal read counts). On a held out test dataset (n = 22), the area under the curve was 0.87 (Fig 3e). Testing gCMS on the data of Farshidfar *et al.* [12], the area under the curve was 0.95 predicting eCMS from gCMS (sensitivity 1, specificity 0.85; Fig. S3f). The gCMS model also allowed us to classify BTCs assessed with the MSK-Impact panel. Using *TP53* and *BAP1* mutations as CMS markers in the data of Song Y. et al. [28], we found an area under the curve of 0.86 (n = 289). Stratifying patients by gCMS for this data (n = 534), we see significantly longer overall survival in CMS-A (median survival of 29.9 months in CMS-A and 23.6 months in CMS-B), adjusting for grade and anatomical location (p = 0.00491, Fig 3f, S3g).

Divergent evolutionary dynamics between subtypes reflect risk factors

Turning to the broader mutational context, we observed a 1.6-fold increase in tumor mutational burden in CMS-B over CMS-A (p = 3.3e-06, Fig 4a). These differences remained significant after excluding highly mutated samples (defined as TMB < 10; fold change = 1.4, p = 6.5e-05), and therefore is not solely explained by the numerical increase in rates of homologous recombination deficiency and microsatellite instability in CMS-B over CMS-A (Fig. 4b). The increased burden was evident for both single basepair substitutions (SBS) (fold change = 1.4, p = 0.0001) and small insertions and deletions (ID) (fold change = 1.8, p = 1.2e-07). CMS-B tumors were also more likely to exhibit extrachromosomal DNA (p = 0.0003; Fig. 4c).

We next compared whether the increased burden could be explained by mutational signatures. *De novo* mutational signature extraction identified 19 SBS, seven double basepair substitution (DBS) and eight ID signatures (Fig. S4). Comparing signature contributions between the subtypes using a linear model, CMS-B tumors showed significant enrichment of SBS1 (clock-like; p = 4.5e-08), ID2 (clock-like; p = 3.0e-06), and SBS2/SBS13 (APOBEC; p = 1.9e-04 and 3.0e-04), whereas CMS-A tumors showed enrichment of ID5

(unknown etiology; p = 2.9e-05; Fig 4d). Further, CMS-specific signatures were found to be a principal component of variance among samples. The second principal component of variance in mutation counts assigned to each signature across samples, explaining 8.4% of the variance, distinguished CMS-A from CMS-B tumors (p = 1.3e-09; Fig. 4e). The first principal component (PC1), explaining 9.8% of the variance, loaded on SBS31 and DBS5, both associated with platinum therapy. As expected, patients who received platinum-based chemotherapy before specimen collection had significantly higher PC1 values (p = 2.4e-05). These results suggest not only significant differences in mutational profiles between subtypes, but that these differences are among the top axes of diversity among samples.

Next, we hypothesized that differences in mutational burdens might reflect tumor evolution. Looking at genetic diversity in subclones, CMS-A showed a significantly higher number of subclones (p = 0.03) and Shannon diversity index (p = 0.008) than CMS-B. Breaking down the distribution of cancer cell fraction, a larger proportion of variants in CMS-A belonged to rare clones (< 0.5 cancer cell fraction), while the majority of variants in CMS-B were clonal or nearly clonal (Fig. 4f). Among clonal mutations, signatures showing a strong CMS-specific bias occurred early that subtype's evolution: relatively more APOBEC and clock-like signatures were earlier than later in CMS-B (p = 0.00024 and p = 6.8e-09, respectively) with a similar finding for ID5 in CMS-A (p = 2.6e-05). Consistent with kyklonic hypermutation [29], APOBEC SBS signatures were over-represented near extrachromosomal DNA in CMS-B ($p < 1e{-}16$). Cases with *MDM2* extrachromosomal DNA had the highest rates of over-representation of SBS3/12 in the cohort. These results suggest APOBEC may be important in early stages of CMS-B.

The divergent mutational landscapes of the CMS suggest differences in early oncogenesis. Therefore, we evaluated associations between each CMS and established BTC risk factors. In our cohort, primary sclerosing cholangitis was significantly more common in CMS-B than -A (p = 0.016; Fig. 4h) . In the cohort of Jusakul *et al.* [11], liver fluke associated tumours were overrepresented in CMS-B (p =1.8e-10; Fig. 4i), while in the Song Y. *et al.* [28] cohort, we found significantly higher rates of cholecystitis in CMS-B (p = 0.0029) and a trend towards higher rates of hepatic steatosis in CMS-A (p = 0.067; Fig. 4j). In the partially overlapping dataset from Boerner *et al.* [14], the association between steatosis and CMS-A was significant (p = 0.023). Together, these differences suggest that CMS-B is associated with chronic inflammatory biliary conditions, which may lead to increased APOBEC mutations and result in higher mutational burdens.

Gene expression profiles of the subtypes reflect different programs

We next assessed how CMS relates to primary anatomic sites. In our cohort using the eCMS classifier, CMS-A occurred exclusively in intrahepatic cholangiocarcinoma or HCC-CCA, while CMS-B spanned the biliary tract (Fig. 5a). 68% (n = 49) of intrahepatic cholangiocarcinomas were CMS-A. We observed similar tissue associations with the gCMS classifier in the dataset from Song Y. *et al.* 2024 [28] (Fig. 5b) . Using the cohort of Jolissaint *et al.* 2021 [30], the gCMS revealed enrichment of large duct intrahepatic cholangiocarcinoma within CMS-B (n = 12, p = 0.01), while the small duct histopathology was significantly enriched in CMS-A (n = 117, p = 0.01; Fig. 5c). The site of biopsy, from metastases with RNA (n = 21), did not affect the association between CMS and the primary site. Notably, all 17 liver metastases from extrahepatic primaries belonged to CMS-B.

Similarly, analyzing 11 primary and metastatic samples paired by donor, all pairs exhibited concordant eCMS despite having different biopsy sites.

Embryologically, small intrahepatic bile ducts derive from a hepatoblast progenitor, whereas extrahepatic ducts share origins with gallbladder and pancreas [31]. To further explore this association, we returned to the whole-transcriptome sequencing data. Differential gene expression showed CMS-B tumors had higher *MUC2, CDX2*, and *KRT20* (CK20) expression. These markers define intraductal papillary neoplasms of large and extrahepatic bile ducts [32,33]. In contrast, CMS-A tumors showed strong and significant increased expression for canonical cholangiocyte markers (*DCDC2*, *BICC1*, *ONECUT1,* and *CFTR*), as well as hepatocyte markers (*ALB*) and hepatoblast markers (*HHEX*, *PROX1*; Fig. 5d). These markers may point to origins from different cholangiocyte populations [34] .

Because bulk transcriptome data collapses expression profiles from multiple cell types, we next performed virtual microdissection using non-negative matrix factorization to separate expression signatures. Optimizing for tumor-associated signatures (see *Methods*), the data were best explained by eight signatures (k) across 4,000 genes (Fig. S5a). Gene set enrichment analysis found two expression signatures (signatures 3 and 4) were significantly enriched for the CMS-B markers identified by the top scoring pair algorithm (normalized enrichment scores = 2.0 and 2.0, FDR = 3e-05 and 2e-05, respectively; Fig. 5e). Both of these signatures were enriched for scRNA markers of common bile duct epithelium [35] (FDR = 2.7e-07 and 1.4e-09) and genes corresponding to rectal goblet cell proteins from Human Protein Atlas[36] were significantly over-represented (FDR = 6.6e-17 and 1.6e-12). Goblet cells can appear in areas of the bile ducts and gallbladder associated with intestinal metaplasia [37]. With the top scoring pair markers for CMS-A, gene set enrichment analysis found one significantly enriched signature: 8 (normalized enrichment scores = 2.2, FDR = 2.7e-09). Gene set enrichment analysis of scRNA markers found enrichment of intrahepatic duct markers (FDR = 9e-10), and, of Human Protein Atlas markers, we found overrepresentation of kidney proximal tubule genes (FDR = 3.7e-05). The molecular connection between proximal tubules of the kidney and small ducts of the biliary tree is unclear, but intriguing given their similar roles in sensing and adjusting osmotic pressure [38]. Signatures 2 and 6 weights were significantly correlated with the composition of multiple immune cell lineages (Fig. S5b). Together, tumor markers and tissue microdissection suggest the CMS may be associated with different types of cholangiocytes.

Using data from Song G. et al. [16], we explored tumor RNA at single cell resolution. We first inferred aneuploidy for each cell, a marker for malignancy [39]. Examining the inferred copy number profiles, we noticed a strong similarity with the copy number alteration profiles characteristic of each CMS based on our previous bulk DNA sequencing (Fig. S5c, d), leading us to classify individual tumour cells using the gCMS model on inferred copy number profiles. On average, 82% of aneuploid cells within tumors from the same patient (n = 11, range 53-100%) were predicted to match the consensus (i.e. most common) gCMS label for that tumour. Next, we subset tumors for aneuploid cells from each patient and inferred eCMS from pseudobulk of aneuploid cells. We found 100% concordance between consensus single-cell gCMS and pseudobulk eCMS for individual tumours. While eCMS was unreliable at single cell resolution due to high rates of gene drop-out for any of the 16 top scoring pairs, we were able to project the richer gene sets associated with our non-negative matrix factorization-derived expression signatures onto single cells (Fig 5f). Tallying cells classified by their top

expression signature [40], we found the majority of aneuploid cells deriving from CMS-B tumors upregulated signature 4 and CMS-A tumor cells upregulated signature 8 (Fig 5g). Quantifiably similar conclusions could be made with the scRNA datasets from Shi *et al.* [41] and Zhang *et al.* [42] (Fig S5e, f), suggesting both the eCMS and gCMS models converge down to single cell resolution and represent fundamentally different tumour cell profiles.

Discussion

Molecular subtyping has stratified many cancers into biologically and clinically meaningful groups [43–45]. However, in BTC, classification has been anchored to anatomic origin or mutations in specific genes [8]. Harnessing the broader genomic and transcriptomic diversity of BTC for subtyping may uncover novel biology, improve prognostic stratification, and guide the design of new treatment strategies.

Unifying international studies of RNA sequencing from cell lines [15], tumor single cells [16], and bulk tissue [17], as well as genome, methylome [11], and proteome [23] profiling, we identified two consensus molecular subtypes of BTC: CMS-A and -B. We next studied our cohort of BTC with whole-genome and transcriptome sequencing, the largest to date, revealing fundamental differences in the subtypes across clinical, genomic, and transcriptomic features. In CMS-B, high mutation rates arising from APOBEC, possibly associated with inflammatory biliary conditions, led to a diversity of genetic drivers, as well as frequent extrachromosomal DNA. Extrachromosomal DNA provided a mechanism for oncogene activation, exemplified by amplification and overexpression of *MDM2*, and is a potential drug target[46]. We also identified CMS-B drivers newly associated with BTC, including *EPHA2.* CMS-A had a distinct profile: tumor cells exhibited expression patterns similar to intrahepatic cholangiocytes with hepatocyte programs, lower mutational burdens, and recurrent large-scale copy number alterations without well-characterized drivers. Interrogating these regions revealed underappreciated putative tumor suppressors in BTC, including *ERRFI1*, which has potential therapeutic implications[47].

BTC-CMS presents a novel biomarker with strong prognostic clinical implications. CMS-B had worse outcomes after surgery in external validations in a Chinese cohort [23] and an American cohort [28]. Furthermore, we found that the BTC-CMS was prognostic in our cohort of advanced stage patients receiving first-line palliative systemic therapy, including immunotherapy, a setting in which transcriptomic classifiers have not been previously evaluated. In contrast, anatomic location was not prognostic after multivariable adjustment. Importantly, BTC-CMS can be derived from a small set of RNA transcripts for the expression-based model or from copy-number loci derived from typical clinical DNA sequencing panels. Therefore, the BTC-CMS could be readily adopted as a prognostic biomarker, and prospective trials could evaluate its potential to guide therapy.

Just as rRNA phylogenies redefined microbial taxonomy and the tree of life [48], molecular data can refine hierarchies in tumor classification. The BTC-CMS provides an organizing framework beyond anatomy that balances the trade-off between distinguishing features of meaningful subtypes [49] while grouping similar tumors to maintain statistical power in clinical trials for these rare cancers [50]. We anticipate that larger cohorts

will reveal further substratification within the subtypes, and additional stratification exists within the tumor microenvironment.

In conclusion, our data provide the first comprehensive integration of multi-omic and clinical evidence to define consensus molecular subtypes for BTC. We offer a coherent framework with direct implications for prognostic stratification. By bridging previously disparate observations and infusing the field with the largest new genomics dataset in a decade, we lay the groundwork for a new era of precision oncology in a historically heterogeneous disease.

References

1. Foà, R. & Chiaretti, S. Philadelphia Chromosome–Positive Acute Lymphoblastic Leukemia. *N. Engl. J. Med.* **386**, 2399–2411 (2022).
2. Mellinghoff, I. K. *et al.* Vorasidenib in IDH1- or IDH2-Mutant Low-Grade Glioma. *N. Engl. J. Med.* **389**, 589–601 (2023).
3. André, T. *et al.* Pembrolizumab in Microsatellite-Instability–High Advanced Colorectal Cancer. *N. Engl. J. Med.* **383**, 2207–2218 (2020).
4. Knox, J. J. *et al.* Whole genome and transcriptome profiling in advanced pancreatic cancer patients on the COMPASS trial. *Nat. Commun.* **16**, 5919 (2025).
5. Stahler, A. *et al.* Individual Patient Data Meta-Analysis of Consensus Molecular Subtypes as Biomarkers of First-Line Treatment in *RAS* Wild-Type Metastatic Colorectal Cancer. *J. Clin. Oncol.* JCO-25-00596 (2025) doi:10.1200/JCO-25-00596.
6. Vogel, A. *et al.* Biliary tract cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. *Ann. Oncol.* **34**, 127–140 (2023).
7. Banales, J. M. *et al.* Cholangiocarcinoma 2020: the next horizon in mechanisms and management. *Nat. Rev. Gastroenterol. Hepatol.* **17**, 557–588 (2020).
8. Beaudry, F. E. G. *et al.* Molecular Tumor Board-Guided Targeted Treatments for Biliary Tract Cancers in a Publicly Funded Healthcare System. *Curr. Oncol.* **32**, 80 (2025).
9. Kendre, G. *et al.* Charting co-mutation patterns associated with actionable drivers in intrahepatic cholangiocarcinoma. *J. Hepatol.* **78**, 614–626 (2023).
10. Mody, K. *et al.* Clinical, Genomic, and Transcriptomic Data Profiling of Biliary Tract Cancer Reveals Subtype-Specific Immune Signatures. *JCO Precis. Oncol.* e2100510 (2022) doi:10.1200/PO.21.00510.
11. Jusakul, A. *et al.* Whole-Genome and Epigenomic Landscapes of Etiologically Distinct Subtypes of Cholangiocarcinoma. *Cancer Discov.* **7**, 1116–1135 (2017).
12. Farshidfar, F. *et al.* Integrative Genomic Analysis of Cholangiocarcinoma Identifies Distinct IDH-Mutant Molecular Profiles. *Cell Rep.* **18**, 2780–2794 (2017).
13. Martin-Serrano, M. A. *et al.* Novel microenvironment-based classification of intrahepatic cholangiocarcinoma with therapeutic implications. *Gut* **72**, 736–748 (2023).
14. Boerner, T. *et al.* Genetic Determinants of Outcome in Intrahepatic Cholangiocarcinoma. *Hepatol. Baltim. Md* **74**, 1429–1444 (2021).
15. Vijay, V. *et al.* Generation of a biliary tract cancer cell line atlas identifies molecular subtypes and therapeutic targets. *Cancer Discov.* https://doi.org/10.1158/2159-8290.CD-24-1383 (2025) doi:10.1158/2159-8290.CD-24-1383.
16. Song, G. *et al.* Single-cell transcriptomic analysis suggests two molecularly distinct subtypes of intrahepatic cholangiocarcinoma. *Nat. Commun.* **13**, 1642 (2022).
17. O'Rourke, C. J. *et al.* Molecular portraits of patients with intrahepatic cholangiocarcinoma who diverge as rapid progressors or long survivors on chemotherapy. *Gut* gutjnl-2023-330748 (2023) doi:10.1136/gutjnl-2023-330748.
18. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385.e18 (2018).
19. Kinnersley, B. *et al.* Analysis of 10,478 cancer genomes identifies candidate driver genes and opportunities for precision oncology. *Nat. Genet.* **56**, 1868–1877 (2024).
20. Holzapfel, N. *et al.* Whole-genome sequencing of 20 cholangiocarcinoma cases reveals unique profiles in patients with cirrhosis and primary sclerosing cholangitis. *J. Gastrointest. Oncol.* **14**, 379–389 (2023).

21. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21**, 1350–1356 (2015).
22. Andersen, J. B. *et al.* Genomic and Genetic Characterization of Cholangiocarcinoma Identifies Therapeutic Targets for Tyrosine Kinase Inhibitors. *Gastroenterology* **142**, 1021-1031.e15 (2012).
23. Dong, L. *et al.* Proteogenomic characterization identifies clinically relevant subgroups of intrahepatic cholangiocarcinoma. *Cancer Cell* **40**, 70-87.e15 (2022).
24. Knudson, A. G. Mutation and Cancer: Statistical Study of Retinoblastoma. *Proc. Natl. Acad. Sci.* **68**, 820–823 (1971).
25. Rheinbay, E. *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, 102–111 (2020).
26. Imielinski, M., Guo, G. & Meyerson, M. Insertions and Deletions Target Lineage-Defining Genes in Human Cancers. *Cell* **168**, 460-472.e14 (2017).
27. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
28. Song, Y. *et al.* A Novel Approach to Quantify Heterogeneity of Intrahepatic Cholangiocarcinoma: The Hidden-Genome Classifier. *Clin. Cancer Res.* **30**, 3499–3511 (2024).
29. Bergstrom, E. N. *et al.* Mapping clustered mutations in cancer reveals APOBEC3 mutagenesis of ecDNA. *Nature* **602**, 510–517 (2022).
30. Jolissaint, J. S. *et al.* Intrahepatic Cholangiocarcinoma with Lymph Node Metastasis: Treatment-Related Outcomes and the Role of Tumor Genomics in Patient Selection. *Clin. Cancer Res.* **27**, 4101–4108 (2021).
31. Lotto, J., Stephan, T. L. & Hoodless, P. A. Fetal liver development and implications for liver disease pathogenesis. *Nat. Rev. Gastroenterol. Hepatol.* **20**, 561–581 (2023).
32. Schlitter, A. M. *et al.* Intraductal papillary neoplasms of the bile duct: stepwise progression to carcinoma involves common molecular pathways. *Mod. Pathol.* **27**, 73–86 (2014).
33. Zen, Y. & Akita, M. Neoplastic Progression in Intraductal Papillary Neoplasm of the Bile Duct. *Arch. Pathol. Lab. Med.* **148**, 989–996 (2024).
34. Komuta, M. *et al.* Histological diversity in cholangiocellular carcinoma reflects the different cholangiocyte phenotypes. *Hepatology* **55**, 1876–1888 (2012).
35. Sampaziotis, F. *et al.* Cholangiocyte organoids can repair bile ducts after transplantation in the human liver. *Science* **371**, 839–846 (2021).
36. Thul, P. J. & Lindskog, C. The human protein atlas: A spatial map of the human proteome. *Protein Sci.* **27**, 233–244 (2018).
37. Sakamoto, H. *et al.* A close relationship between intestinal metaplasia and Cdx2 expression in human gallbladders with cholelithiasis. *Hum. Pathol.* **38**, 66–71 (2007).
38. Gunay-Aygun, M. Liver and kidney disease in ciliopathies. *Am. J. Med. Genet. C Semin. Med. Genet.* **151C**, 296–306 (2009).
39. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).
40. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
41. Shi, X. *et al.* Single-cell atlas of diverse immune populations in the advanced biliary tract cancer microenvironment. *Npj Precis. Oncol.* **6**, 58 (2022).
42. Zhang, M. *et al.* Single-cell transcriptomic architecture and intercellular crosstalk of human intrahepatic cholangiocarcinoma. *J. Hepatol.* **73**, 1118–1130 (2020).

43. Ceccarelli, M. *et al.* Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell* **164**, 550–563 (2016).
44. Notta, F. *et al.* A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns. *Nature* **538**, 378–382 (2016).
45. The Cancer Genome Atlas Research Network. Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. *N. Engl. J. Med.* **374**, 135–145 (2016).
46. Tang, J. *et al.* Enhancing transcription–replication conflict targets ecDNA-positive cancers. *Nature* **635**, 210–218 (2024).
47. Borad, M. J. *et al.* Integrated Genomic Characterization Reveals Novel, Therapeutically Relevant Drug Targets in FGFR and EGFR Pathways in Sporadic Intrahepatic Cholangiocarcinoma. *PLoS Genet.* **10**, e1004135 (2014).
48. Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci.* **74**, 5088–5090 (1977).
49. Chakravarty, D. & Solit, D. B. Clinical cancer genomic profiling. *Nat. Rev. Genet.* **22**, 483–501 (2021).
50. Murciano-Goroff, Y. R., Uppal, M., Chen, M., Harada, G. & Schram, A. M. Basket Trials: Past, Present, and Future. *Annu. Rev. Cancer Biol.* **8**, 59–80 (2024).

Methods

Patients and cohort

Patients included in this study contributed to the LeGresley Biliary Registry. This registry prospectively recruits consentable adults over 18 years of age, diagnosed with BTCs at the University Health Network, including individuals with advanced-stage disease. Eligible patients were identified for recruitment through a systematic review of clinic lists. Demographic and clinical data were prospectively collected from the time of enrollment.

Ethics approval and consent to participate

Ethics approval was granted by the University Health Network institutional review board (REB# 24-5412, approved September 11, 2025), and informed consent was obtained from all subjects in this study. All processes were conducted in accordance with the Declaration of Helsinki.

Sample processing and sequencing

A unique tumor sample, from the primary lesion or a metastatic site, with paired normal tissue was included per patient. Samples then underwent whole genome sequencing, prioritizing fresh frozen laser capture microdissected samples, and then whole transcriptome sequencing if sufficient tissue remained. Whole-genome and transcriptome libraries were prepared and Illumina sequencing was performed by the Ontario Institute of Cancer Research. Sample processing is previously described in Beaudry *et al* [8].

Whole genome sequencing fastq files from each lane were adapter trimmed using *cutadapt* 1.8 [51] and aligned to hg38.p14 using *bwa mem* 0.7.17 [52]. Optical replicates were marked by *Picard* 2.21.4 and lanes collapsed into a single .bam file. Median coverage was estimated with *Picard CollectWgsMetrics,* to a minimum of 80X for tumor samples and 30X for normal samples; additional sequencing was performed when this threshold was not met. For estimating tumor content, B-allele frequencies were derived from germline single nucleotide variants called using GATK's v. 4.1.2 *haplotypecaller* [53], with calls recalibrated and realigned using *Picard*. Read depth was segmented and integrated with B-allele frequencies using *HMMCopy* 0.1.1 and copy number alterations, tumor content and ploidy were called using *Celluloid* v0.11.7 [44]. All *celluloid* solutions were manually reviewed and samples estimated to have less than 30% cellularity were excluded.

RNA reads were mapped to hg38.p14 using *star* 2.7.4a [54]. Fusions were identified using *star-fusion* 1.9.0 and annotated with *mavis* v2.2.9 [55]. Reads were tallied per transcript using *htseq* v2.0.4 [56]. Whole transcriptome samples where the median reads per gene was less than one were excluded. All analyses in R were performed on version 4.5.1 and in python on version 3.10.6.

Expression class network and consensus molecular subtyping

For subtype assignment, transcripts were summed at the gene level using annotations from Ensembl release 100 for all RNA samples with frozen fresh specimens. Count data were normalized using *Deseq2* v 1.48.1 [57] and scaled using *ematAdjust* from *CMScaller* v0.99.2 [58] . Samples were assigned subtypes using Nearest Template

Prediction (*ntp*) with *CMScaller* over 1,000 permutations. Where the false discovery rate of class assignment of a sample was greater than 0.05, the class assignment for that classifier was considered undefined.

To build the molecular classifier network, we included 23 signatures from nine previously published BTC tumor classifiers (Table S2). Signature similarity was assessed using a Jaccard similarity matrix. Classifiers (nodes) were removed where (1) a classifier had a lack of conclusive assignment in >50% of cases or (2) was connected to the rest of the network by less than 2% of samples. Additional signatures were removed in manual optimization of the network's inflation factor by visualizing modularity and silhouette scores, evaluating values between 1.5 and 20 by steps of 0.5 (Fig. S2c). Network clusters were identified with the Markov cluster algorithm *mcl* v 1.0 [21].

To train a top-scoring pair algorithm, we restricted the training set to samples that were assigned to the same CMS across every classifier. The algorithm was trained on 80% of the data with the *SWAP.KTSP.Train* function of *switchBox* v 1.44.0 [59] in the range from 10:16 genes pairs and tested on the remaining 20%. Of the remaining samples where individual classifiers disagreed on the consensus cluster assignment, we assigned CMS when >12/16 of gene pairs agreed but did not assign a class where less than 12 of 16 gene pairs were aligned as this parameter space was undefined in the training and test sets (Fig. S2d).

For external testing, RNA data from Dong *et al* [23] was extracted from Table S3 and proteins from Table S5 of that manuscript. RNA beadchip data from Jusakul et al [11] was downloaded from cbioportal.org and normalized with *lumi* v 2.60.0 [60], and methylation data was downloaded from GEO Series GSE89803 and normalized with *minfi* v 1.54.1 [61]. Principal component analysis was performed with *FactoMineR* v. 2.12. Survival differences between the CMS was evaluated in a cox proportional hazard model including overall stage at diagnosis and anatomical location of the primary when appropriate (samples from Dong *et al.* were all intrahepatic) using *coxph* from *survival* v 3.8.3.

Genomic variant calling

Small substitutions (< 1 read in length) including single basepair substitutions (SBS), double basepair substitutions (DBS) and insertion/deletions (in/dels) were called using GATK *mutect2* and *strelka2* 2.9.10 [62] and combined on consensus of both callers. Structural variants were called using *SVaBa* 1.2.0 [63], *delly2* 1.0.3 [64] and *manta* 1.6.0 [65], and combined on consensus of at least two of the three callers.

We used *ActiveDriverWGS* v 1.2.1 [66] to look at statistically recurrent small mutations while controlling for variation in mutation rate across the genome. To look more directly for signals of selection in protein-coding genes, we also used *dNdScv* v 0.0.1.0 [67]. Promoter regions were tested exclusively with *ActiveDriverWGS*. All significantly recurrent hits were manually reviewed in integratedGenomeViewer [68] to exclude germline variants. We performed a Fisher's exact test querying whether any genes were enriched for loss-of-heterozygosity when also hit by a small substitution.

Expression quantitative trait loci were identified using a linear regression model for transcript per million as a function of the presence of a structural variant within 5 kb of the gene of interest, using p-values adjusted with Benjamini-Hochberg correction. Breakpoints were tested for statistical recurrence using *FishHook*

v 0.1 [26]. Extrachromosomal DNA was inferred using *AmpliconArchitect* 1.3.3 [69], downsampling to 8 reads (--downsample 8) and copy number gain threshold to 8 (--cngain 8); runs where a solution was not reached in 20 days were excluded from ecDNA analyses ($n = 6$). Driver co-occurrence significance was corrected for variant frequency using *Rediscover* v 0.3.2 [70]. *GISTIC* v2.0 [27] was used for testing recurrent copy number alterations and converted into gene-level scores using *CNtools* v 1.64.0. To isolate the functional consequences of copy number alterations, we filtered genes in recurrently altered regions for a significant correlation between relative copy number and gene expression measured in transcripts per million.

To train a copy number-based CMS classifier, the top 10 arms loading onto the first principal component of inter-sample arm-level relative copy number alteration variance were included in a generalized linear model and arms were dropped step-wise to minimize the number of factors in the model using the *step* function from *stats* v 4.5.1. Given the use of relative copy number, incorporating the mean variant allele frequency of exonic mutations (as a proxy for tumor cellularity) significantly improved the fit of the model. For external validation of the gCMS classifier, data from Farshidfar *et al.* [12] and Song Y. *et al* [28] were downloaded from cbioportal.org .

Mutational landscape

Tumour mutational burden was calculated as the sum of all small substitutions divided by genome size (in megabases). Microsatellite instability was called using *msisensor-pro* 1.2.0 [71] thresholding MSI-H as greater than 15% of sites showing instability. Candidate homologous recombination deficiency cases were identified using *HRDetect* from *signature.tools.lib* 2.2.0 [72,73] with a cutoff of 90% probability of HRD and reviewed manually. Mutational signature analysis was performed *de novo* using SigProfilerExtractor v 1.2.2 [74], with standard parameters, and decomposition into COSMIC v3.4 reference signatures performed using non-negative least squares. Signature analysis was performed separately for SBS, DBS and In/Dels using somatic variant calls. eCMS signature enrichment was performed using a binomial linear regression model [75], with the response variable as a binary variable representing the appearance of a base assigned to that signature in the given sample, filtering for signatures present in more than 30% of samples.

Clonality was called using *pyclone-IV v* 0.1.6 [76]. Mutations were timed using the mutationTime function with n.boot = 10 and rho = 0 in *MutationTimer* v 1.0.2 [77]. Mutations where both clonality callers disagreed were discarded from clonality analyses. Signature timing was tested by comparing between clonal early and clonal late the distributions in relative exposures in a Wilcoxon rank-sum test. APOBEC enrichment near extrachromosomal DNA was tested at the individual level as enrichment of APOBEC signature SBS2 and SBS13 on chromosomes with extrachromosomal DNA compared to the expectation given equal distribution of APOBEC signatures across chromosomes using a χ^2 test. The cohort level test was performed by summing individual χ^2 values.

Each risk factor association was tested independently by assigning eCMS to samples obtained as described above and frequencies compared between eCMS using a Fisher's exact test. Tests of the anatomical location of primary tumours was performed using the same approach.

Gene expression profiles and virtual microdissection

Differential gene expression between samples from each CMS was evaluated using *DESeq2*. Non-negative matrix factorization (NMF) was performed 200 times (nrun = 200) on RNA read counts per million minimizing the Frobenius norm[78] using *NMF* v 0.28 . To identify malignant components, we normalized component weights, subset to samples where weight was larger than zero for each component, and filtered for components that had a significant correlation with genomics-estimated tumor content greater than 0.2. Optimization of k was performed by finding k where both the average correlation between signature weight and tumor content across components was maximised. NMF weights were converted to pseudo transcript per million values for comparison among samples.

Single cell RNAseq biliary duct cell type markers were downloaded from Table S1 from Sampaziotis *et al.* [35] and single nucleus data from Table S10 from Andrews *et al.* [79] and run on *GSEA* from *clusterProfiler* v 4.16.0 [80]. Human Protein Archive enrichment was performed using *gost* from *gprofiler2* v 0.2.3. *CIBERSORT* correlations were obtained from *immunedeconv* v 2.1.0 [81] deconvolution with "cibersort_abs".

Public single cell RNAseq data analysis

Data from Zhang et al [42] were downloaded from GEO Series GSE138709, from Shi et al [41] from GSE201425 and data from Song G et al [16] were obtained by personnel communication with study authors. Data were normalized using *SCTransform* in *Seurat* v 5.3.0 [82]. We used markers from CellMarker [83] to define cell populations, subsetting to cell populations with epithelial markers or where marker-based cell identity was inappropriate for the context (e.g. acinar cells in the liver) as candidate malignant cell populations. NMF signatures were projected onto single cells by selecting the genes with the top 90th percentile of weights in their respective signatures, and testing for area under the curve-based enrichment with *AUCell_calcAUC* from *AUCell* v 1.30.1 [40]. Copy number alteration profiles were inferred from *infercnv* v 1.24.0 [39] per sample, using immune cell populations as reference and including a denoising step (denoise=TRUE). eCMS call confidence was estimated by resampling 100 subpopulations of 20 cells per tumour sample.

Data availability

All code will be made available on github, and all sequencing data on EGA.

Acknowledgements

This research was supported by the Richard LeGresley Biliary Cancer Research Fund through the Princess Margaret Cancer Foundation. Sequencing was supported by The Terry Fox Research Institute (TFRI) Marathon of Hope Cancer Centres Network (MOHCCN). We thank all patients who contributed to this research.

Conflicts of interest

RG: Received a graduate scholarship from Pfizer and provided consulting or advisory roles for Astrazeneca, Tempus, Eisai, Incyte, Knight Therapeutics, Guardant Health, and Ipsen; **AV**: Consultancy and advisory role: Roche, AstraZeneca, Böhringer- Ingelheim, Ipsen, Incyte, Cogent, EISAI, Zymeworks, Biologix, BMS, Terumo, Elevar, Servier, MSD, Tahio, Jazzpharma, Medivir, Abbvie, Tyra, Falk, Janssen, Lilly.

51. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).
52. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
53. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. Preprint at https://doi.org/10.1101/201178 (2017).
54. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
55. Reisle, C. *et al.* MAVIS: merging, annotation, validation, and illustration of structural variants. *Bioinformatics* **35**, 515–517 (2019).
56. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
57. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
58. Eide, P. W., Bruun, J., Lothe, R. A. & Sveen, A. CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models. *Sci. Rep.* **7**, 16618 (2017).
59. Afsari, B., Fertig, E. J., Geman, D. & Marchionni, L. switchBox: an R package for k–Top Scoring Pairs classifier development. *Bioinformatics* **31**, 273–274 (2015).
60. Du, P., Kibbe, W. A. & Lin, S. M. *lumi* : a pipeline for processing Illumina microarray. *Bioinformatics* **24**, 1547–1548 (2008).
61. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
62. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
63. Wala, J. A. *et al.* SvABA: genome-wide detection of structural variants and indels by local assembly. *Genome Res.* **28**, 581–591 (2018).
64. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
65. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
66. Zhu, H. *et al.* Candidate Cancer Driver Mutations in Distal Regulatory Elements and Long-Range Chromatin Interaction Networks. *Mol. Cell* **77**, 1307-1321.e10 (2020).
67. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041.e21 (2017).
68. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
69. Deshpande, V. *et al.* Exploring the landscape of focal amplifications in cancer using AmpliconArchitect.

Nat. Commun. **10**, 392 (2019).
70. Ferrer-Bonsoms, J. A., Jareno, L. & Rubio, A. Rediscover: an R package to identify mutually exclusive mutations. *Bioinformatics* **38**, 844–845 (2022).
71. Jia, P. *et al.* MSIsensor-Pro: Fast, Accurate, and Matched-Normal-Sample-Free Detection of Microsatellite Instability. *Genomics Proteomics Bioinformatics* **18**, 65–71 (2020).
72. Davies, H. *et al.* HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.* **23**, 517–525 (2017).
73. Degasperi, A. *et al.* A practical framework and online tool for mutational signature analyses show intertissue variation and driver dependencies. *Nat. Cancer* **1**, 249–263 (2020).
74. Islam, S. M. A. *et al.* Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genomics* **2**, 100179 (2022).
75. Díaz-Gay, M. *et al.* Geographic and age variations in mutational processes in colorectal cancer. *Nature* **643**, 230–240 (2025).
76. Gillis, S. & Roth, A. PyClone-VI: scalable inference of clonal population structures using whole genome data. *BMC Bioinformatics* **21**, 571 (2020).
77. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
78. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).
79. Andrews, T. S. *et al.* Single-cell, single-nucleus, and spatial transcriptomics characterization of the immunological landscape in the healthy and PSC human liver. *J. Hepatol.* **80**, 730–743 (2024).
80. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS J. Integr. Biol.* **16**, 284–287 (2012).
81. Sturm, G., Finotello, F. & List, M. Immunedeconv: An R Package for Unified Access to Computational Methods for Estimating Immune Cell Fractions from Bulk RNA-Sequencing Data. in *Bioinformatics for Cancer Immunotherapy* (ed. Boegel, S.) vol. 2120 223–232 (Springer US, New York, NY, 2020).
82. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* **42**, 293–304 (2024).
83. Hu, C. *et al.* CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Res.* **51**, D870–D876 (2023).

Figures

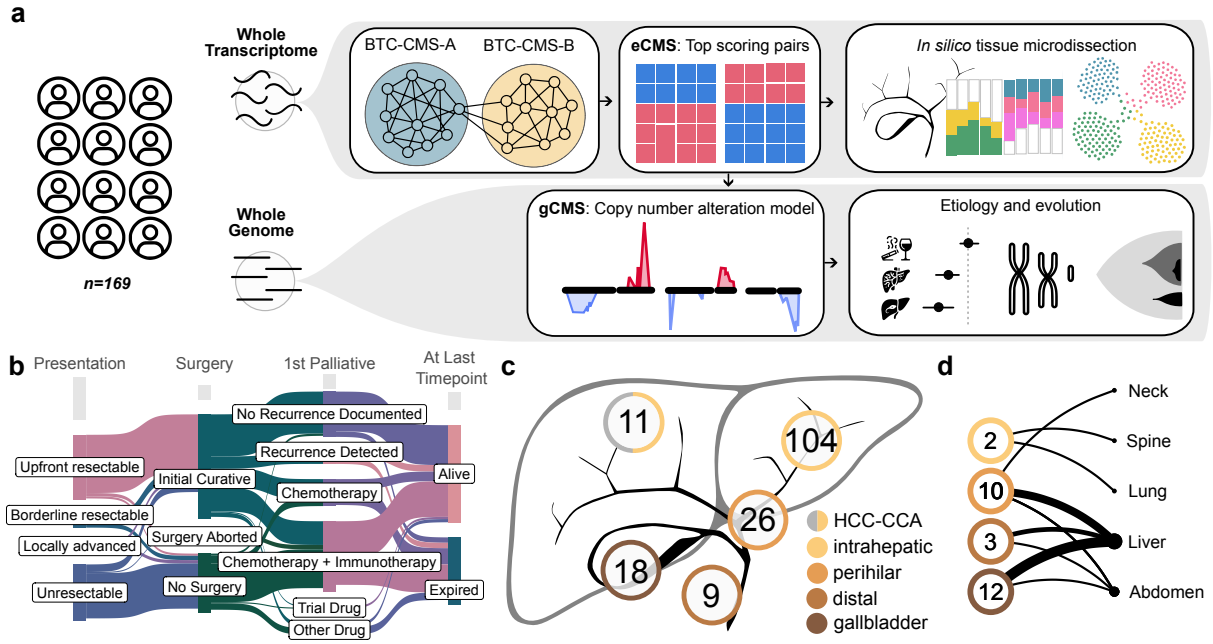

Figure 1

Fig. 1 | Study and cohort overview.

a, study schema. **b**, proportions of patients proceeding through different clinical trajectories. **c**, counts of primary anatomical tumor site. **d**, counts of biopsy sites of metastatic samples. WTS, whole transcriptome sequencing. WGS, whole genome sequencing. BTC, biliary tract cancer. CMS, consensus molecular subtype. HCC-CCA, mixed hepatocellular cholangiocarcinoma. eCMS: expression-based consensus molecular subtypes; gCMS: genetics-based consensus molecule subtypes.

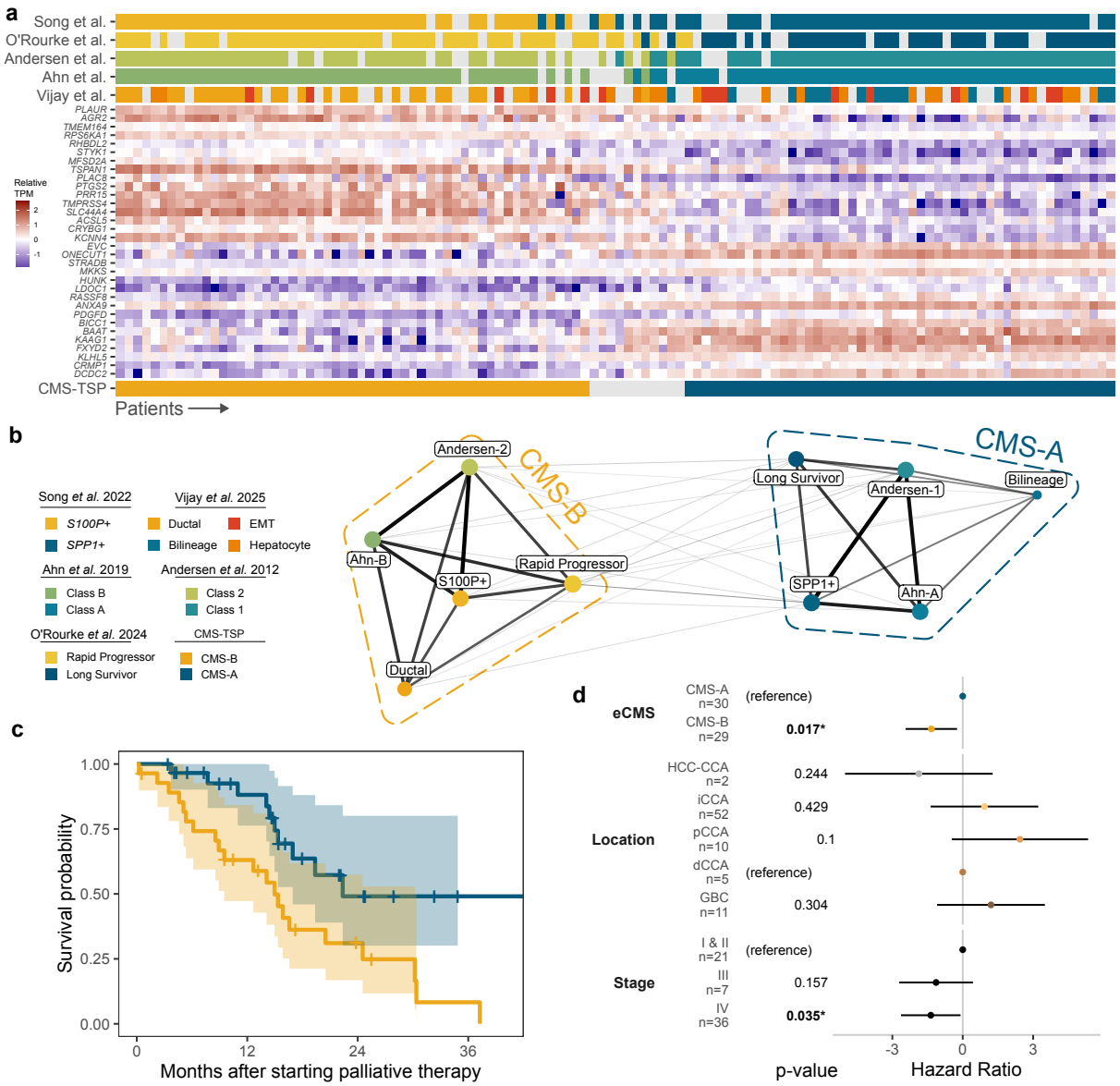
a
Song et al.
O'Rourke et al.
Andersen et al.
Ahn et al.
Vijay et al.
Relative TPM
CMS-TSP
Patients
b
Song et al. 2022
S100P+
SPP1+
Vijay et al. 2025
Ductal
EMT
Bilineage
Hepatocyte
Ahn et al. 2019
Class B
Class A
Andersen et al. 2012
Class 2
Class 1
O'Rourke et al. 2024
Rapid Progressor
Long Survivor
CMS-TSP
CMS-B
CMS-A
Andersen-2
CMS-B
Ahn-B
S100P+
Rapid Progressor
Ductal
CMS-A
Long Survivor
Andersen-1
Bilineage
SPP1+
Ahn-A
c
Survival probability
Months after starting palliative therapy
d
eCMS
CMS-A n=30 (reference)
CMS-B n=29 0.017*
Location
HCC-CCA n=2 0.244
iCCA n=52 0.429
pCCA n=10 0.1
dCCA n=5 (reference)
GBC n=11 0.304
Stage
I & II n=21 (reference)
III n=7 0.157
IV n=36 0.035*
p-value
Hazard Ratio

Figure 2

Fig. 2 | Consensus molecular subtypes of biliary tract cancer are prognostic.

a, assignment of 118 fresh frozen whole transcriptome samples to 6 published classifiers, as well as a heatmap of the 16 genes in the classifier and the derived CMS-based top-scoring pair classifier. **b**, assignment similarity network between the CMS. **c**, overall survival after start of first-line palliative treatment by expression CMS in the present cohort. **d**, cox proportional hazard across expression CMS, primary tumour location, and tumour stage. eCMS, expression consensus molecular subtype. CCA: cholangiocarcinoma; mixed hepatocellular cholangiocarcinoma (HCC-CCA), iCCA: intrahepatic CCA; pCCA: perihilar CCA, dCCA: distal CCA, GBC: gallbladder cancer. TPM, transcripts per million. TSP, top scoring pairs. CMS, consensus molecular subtype.

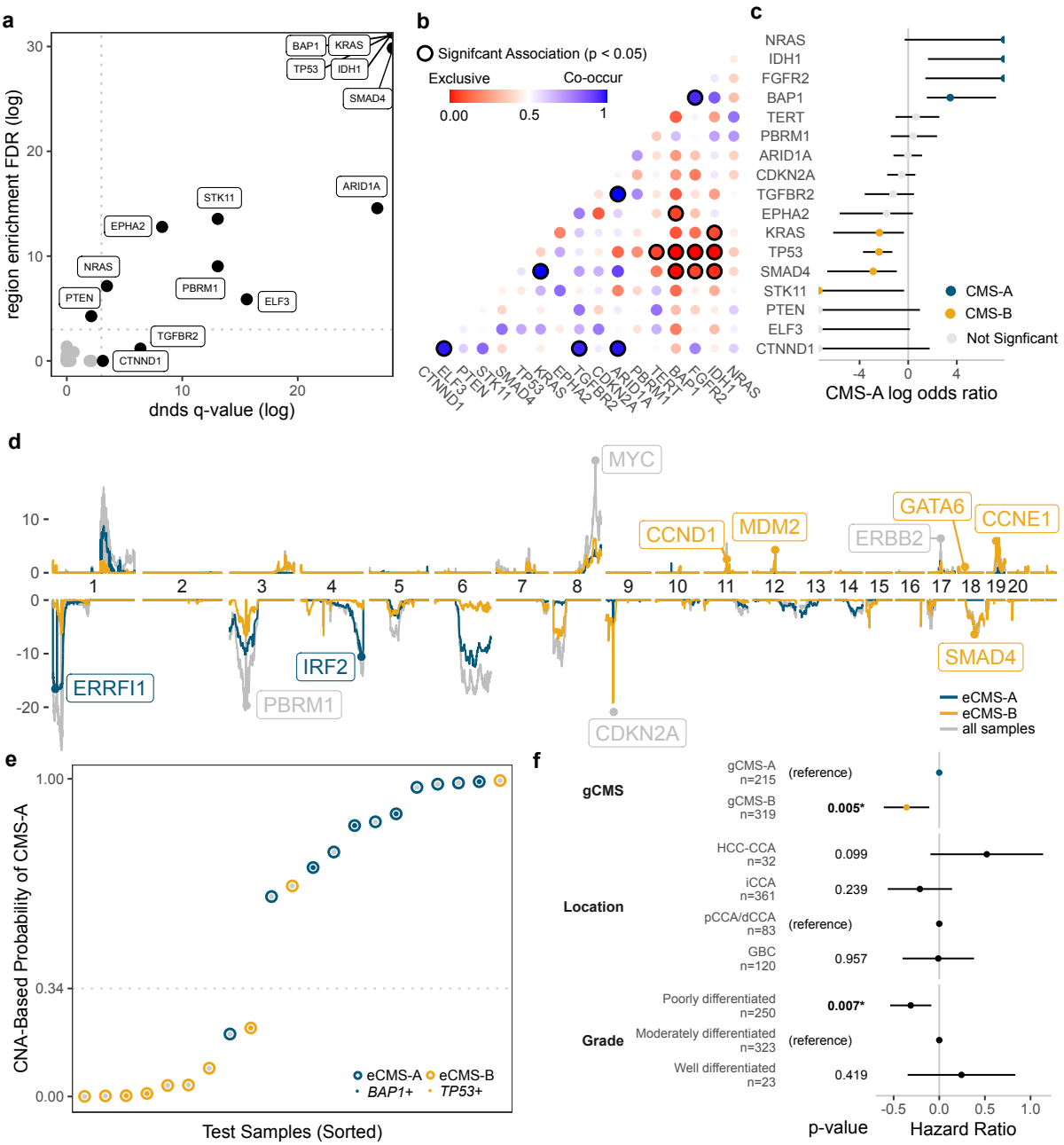
a
region enrichment FDR (log)
dnds q-value (log)
BAP1
KRAS
TP53
IDH1
SMAD4
ARID1A
STK11
EPHA2
NRAS
PBRM1
ELF3
PTEN
TGFBR2
CTNND1
b
Signifcant Association (p < 0.05)
Exclusive
Co-occur
0.00
0.5
1
c
CMS-A
CMS-B
Not Signficant
CMS-A log odds ratio
d
MYC
CCND1
MDM2
GATA6
CCNE1
ERBB2
SMAD4
ERRFI1
IRF2
PBRM1
CDKN2A
eCMS-A
eCMS-B
all samples
e
CNA-Based Probability of CMS-A
Test Samples (Sorted)
eCMS-A
eCMS-B
BAP1+
TP53+
f
gCMS
gCMS-A n=215 (reference)
gCMS-B n=319 0.005*
Location
HCC-CCA n=32 0.099
iCCA n=361 0.239
pCCA/dCCA n=83 (reference)
GBC n=120 0.957
Grade
Poorly differentiated n=250 0.007*
Moderately differentiated n=323 (reference)
Well differentiated n=23 0.419
p-value
Hazard Ratio

Figure 3

Fig. 3 | Genetic drivers of biliary tract cancer.

a, Statistical recurrence analysis for small mutation drivers for a gene-centric approach (the rate of non-synonymous to synonymous mutations, dn/ds) and gene-agnostic approach (regional enrichment). **b**, Co-occurrence matrix for drivers, with enrichment for co-occurence in blue and mutual exclusion in red, and significant associations circled in black. **c**, Log of the odds ratio for driver mutations by consensus molecular subtype (CMS). **d**, Significance of copy number alteration recurrence analysis using GISTIC, across chromosomes by CMS and using all samples (globally). **e**, Accuracy of copy number alteration generalized linear model at predicting expression CMS (eCMS) and small mutation drivers for a test cohort (n = 24). **f**, Cox proportional hazard for genomic CMS (gCMS), location and grade on overall survival from diagnosis for unresectable disease for 534 samples from Song Y. *et al* 2024 . CCA: cholangiocarcinoma; mixed hepatocellular cholangiocarcinoma (HCC-CCA), iCCA: intrahepatic CCA; pCCA: perihilar CCA, dCCA: distal CCA, GBC: gallbladder cancer. FDR: False discovery rate. CNA: copy number alteration.

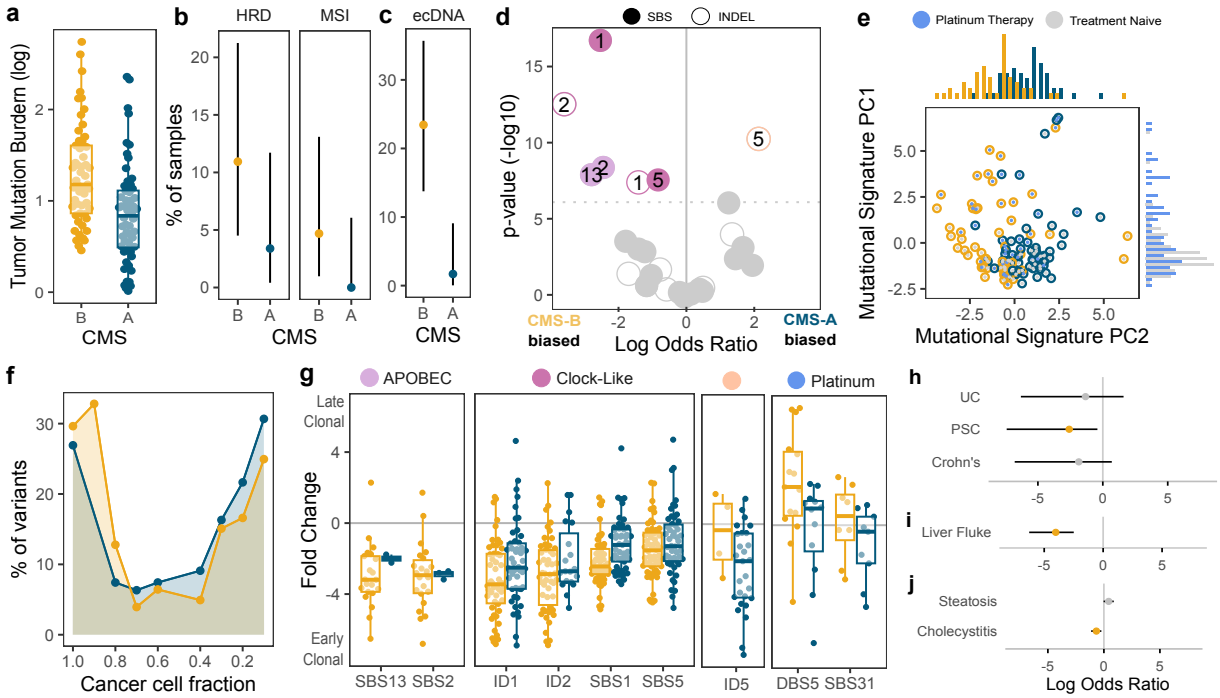
a
Tumor Mutation Burdern (log)
CMS
b
HRD
MSI
% of samples
CMS
c
ecDNA
CMS
d
SBS
INDEL
p-value (-log10)
CMS-B biased
Log Odds Ratio
CMS-A biased
e
Platinum Therapy
Treatment Naive
Mutational Signature PC1
Mutational Signature PC2
f
% of variants
Cancer cell fraction
g
APOBEC
Clock-Like
Platinum
Late Clonal
Early Clonal
Fold Change
SBS13 SBS2
ID1
ID2
SBS1
SBS5
ID5
DBS5 SBS31
h
UC
PSC
Crohn's
i
Liver Fluke
j
Steatosis
Cholecystitis
Log Odds Ratio

Figure 4

Fig. 4 | Etiologies and mutation profiles of biliary tract cancers.

a, Tumor mutational burden difference between consensus molecular subtypes (CMS). **b**, frequency of samples in each CMS that show mutator phenotypes of homologous recombination deficiency (HRD) and microsatellite instability (MSI). **c**, frequency of samples in each CMS that show signatures of extrachromosomal DNA (ecDNA). **d**, average number of mutations per subclone at each cancer cell fraction for each CMS. **f**, association between CMS and mutational signatures for single basepair substitutions and in/dels and their etiologies. **e**, principal component analysis of mutational signatures showing platinum therapy exposed samples along the first principal component and CMS on the second. **f**, average number of mutations per subclone at each cancer cell fraction for each CMS. **g**, relative evolutionary timing of clonal mutations by mutational signature and CMS. **h-j**, association between consensus molecular subtype and biliary tract cancer risk factors: **h**, inflammatory bowel disease with bile duct involvement, **i**, liver fluke, **j**, steatosis and cholecystitis. SBS: single basepair substitutions; DBS: double basepair substitutions; PC: principal component; NHEJ: non-homologous end-joining; UC: ulcerative colitis; PSC: primary sclerosing cholangitis.

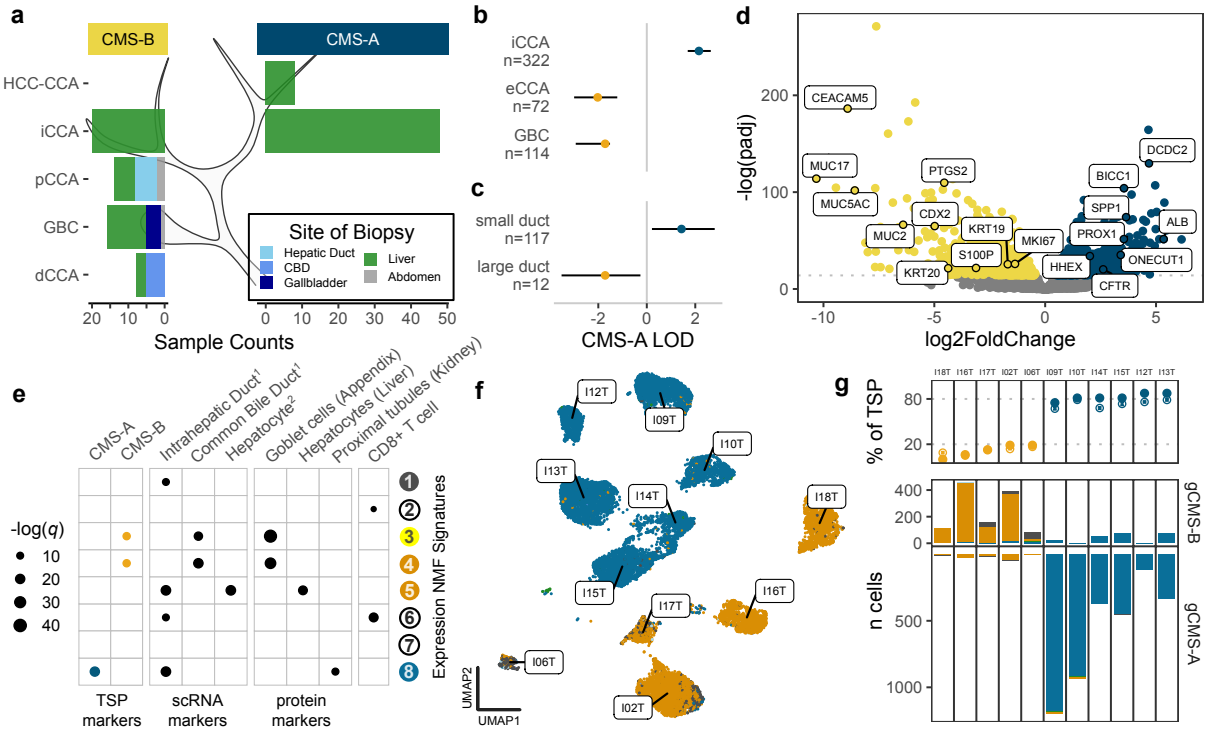
a
CMS-B
CMS-A
HCC-CCA
iCCA
pCCA
GBC
dCCA
Site of Biopsy
Hepatic Duct
CBD
Gallbladder
Liver
Abdomen
Sample Counts
b
iCCA n=322
eCCA n=72
GBC n=114
c
small duct n=117
large duct n=12
CMS-A LOD
d
-log(padj)
log2FoldChange
CEACAM5
MUC17
MUC5AC
PTGS2
CDX2
MUC2
KRT19
S100P
KRT20
MKI67
HHEX
PROX1
SPP1
BICC1
DCDC2
ALB
ONECUT1
CFTR
e
CMS-A
CMS-B
Intrahepatic Duct
Common Bile Duct
Hepatocyte
Goblet cells (Appendix)
Hepatocytes (Liver)
Proximal tubules (Kidney)
CD8+ T cell
-log(q)
Expression NMF Signatures
TSP markers
scRNA markers
protein markers
f
UMAP1
UMAP2
g
% of TSP
n cells
gCMS-B
gCMS-A

Figure 5

Fig. 5 | *In silico* dissection of tissue and cell types for consensus molecular subtypes.

a, Counts of tumors classified by expression consensus molecular subtype (CMS) into anatomical location of primary tumor on the biliary tract, as well as specimen location. **b-c**, Likelihoods for CMS-A associated across anatomical location of primary tumor for **b,** n = 508 samples from Song Y. *et al.* (2024) and **c,** n = 129 samples from Jolissaint *et al.* (2021). **d**, Differential gene expression between subtypes (n = 116). **e**, cell population marker enrichment for expression components from non-negative matrix factorization, including scRNA markers from [1] Sampaziotis *et al.* (2021) and [2] Andrews *et al.* (2024) . **f**, Epithelial cells from single cell RNA from Song G. *et al.* by top expression signature and labelled by tumor. **g**, eCMS per tumor by pseudobulk TSP, and tallies of cells by top NMF component split by single cell genetic CMS (gCMS). TSP: top scoring pair.

Supplementary Figures

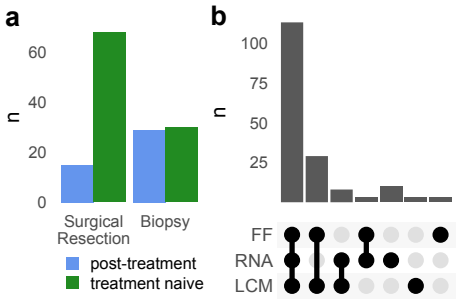

Figure S1

Fig. S1 | Cohort specimens

a, counts of specimens from biopsy compared to surgical resection, by timepoints of treatment naive (green) and post-treatment (blue). **b,** counts of specimens by collection features including paired RNA (WTS), laser capture microdissection (LCM), and fresh frozen (FF) as opposed to formalin-fixed paraffin embedded (FFPE); most samples are LCM fresh frozen samples with paired RNAseq (WTS).

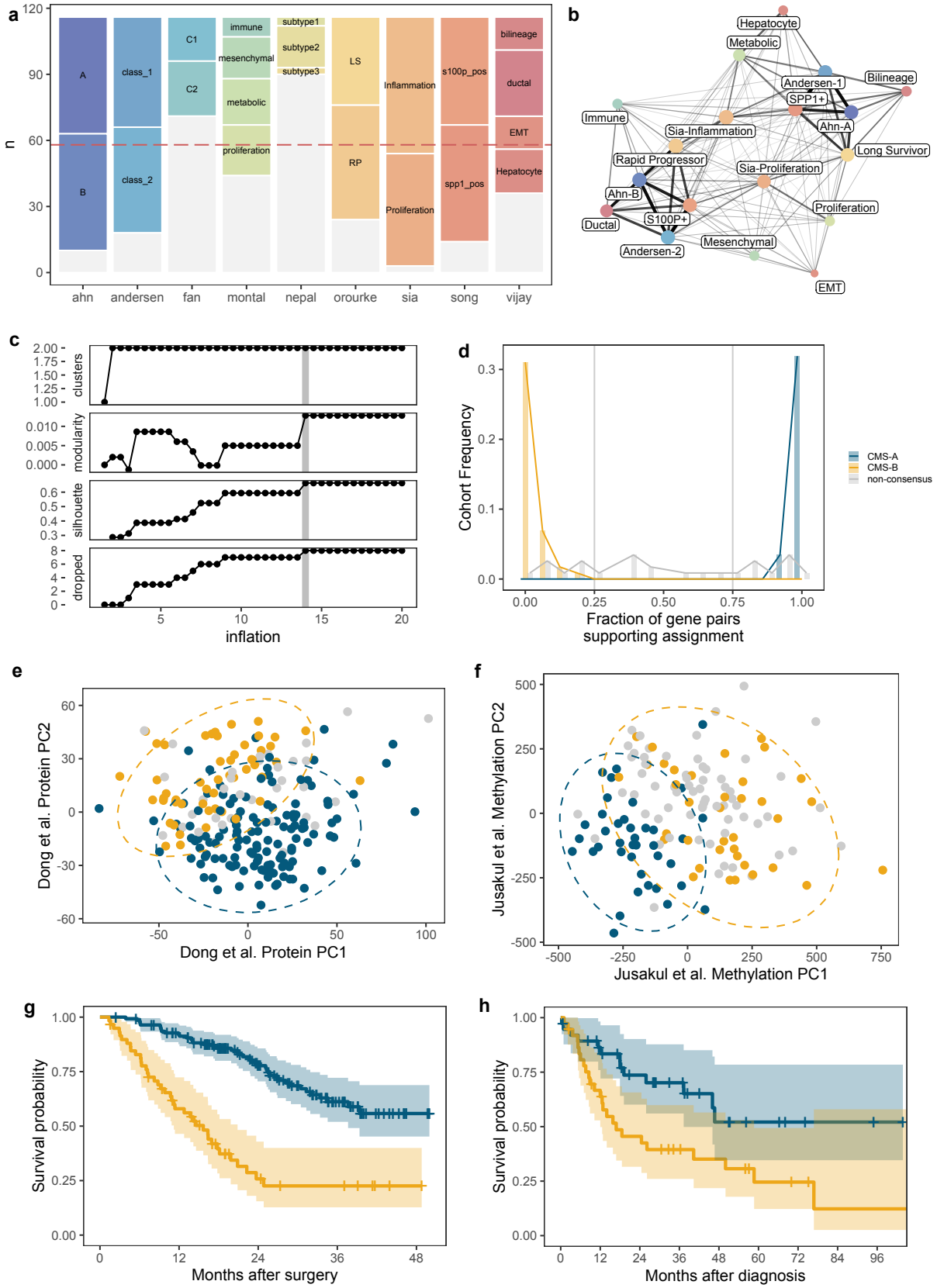
a
n
A
B
class_1
class_2
C1
C2
immune
mesenchymal
metabolic
proliferation
subtype1
subtype2
subtype3
LS
RP
Inflammation
Proliferation
s100p_pos
spp1_pos
bilineage
ductal
EMT
Hepatocyte
ahn
andersen
fan
montal
nepal
orourke
sia
song
vijay
b
Hepatocyte
Metabolic
Bilineage
Andersen-1
SPP1+
Immune
Sia-Inflammation
Ahn-A
Long Survivor
Rapid Progressor
Sia-Proliferation
Ahn-B
Proliferation
Ductal
S100P+
Mesenchymal
Andersen-2
EMT
c
clusters
modularity
silhouette
dropped
inflation
d
Cohort Frequency
Fraction of gene pairs supporting assignment
CMS-A
CMS-B
non-consensus
e
Dong et al. Protein PC2
Dong et al. Protein PC1
f
Jusakul et al. Methylation PC2
Jusakul et al. Methylation PC1
g
Survival probability
Months after surgery
h
Survival probability
Months after diagnosis

Figure S2

Fig. S2 | eCMS training cohort and external validation

a, assignments of 116 fresh frozen RNA samples to respective classes from nine previously published classifiers. Assignment fit threshold (dashed line) was set for classifiers to have greater than 50% of cases assigned to a class (FDR > 0.05). **b,** expression classifier network before optimization. **c,** four parameters used for manual network optimization across inflation factors: number of clusters, modularity, silhouette score, and number of classes dropping out of the network. The gray band shows the chosen optimal inflation factor. **d,** fraction of pairs in TSP eCMS algorithm supporting eCMS classification for samples classified as CMS-A, CMS-B and samples where classifiers within network clusters disagreed (non-consensus; gray), according to their frequency in the cohort. All classifiers agreed on assigning classes to samples where >75% of pairs agreed (vertical gray lines). **e-f,** principal component analysis results classified by eCMS on, **e,** global proteomic profiling from 262 treatment-naive iCCAs from Dong *et al*. and, **f,** 485k methylation array probes from 75 BTCs from Jusakul *et al*. **g-h.** Kaplan-Meir analysis of survival **g.** after surgery from Dong *et al.* data and **h.** Jusakul *et al* data.

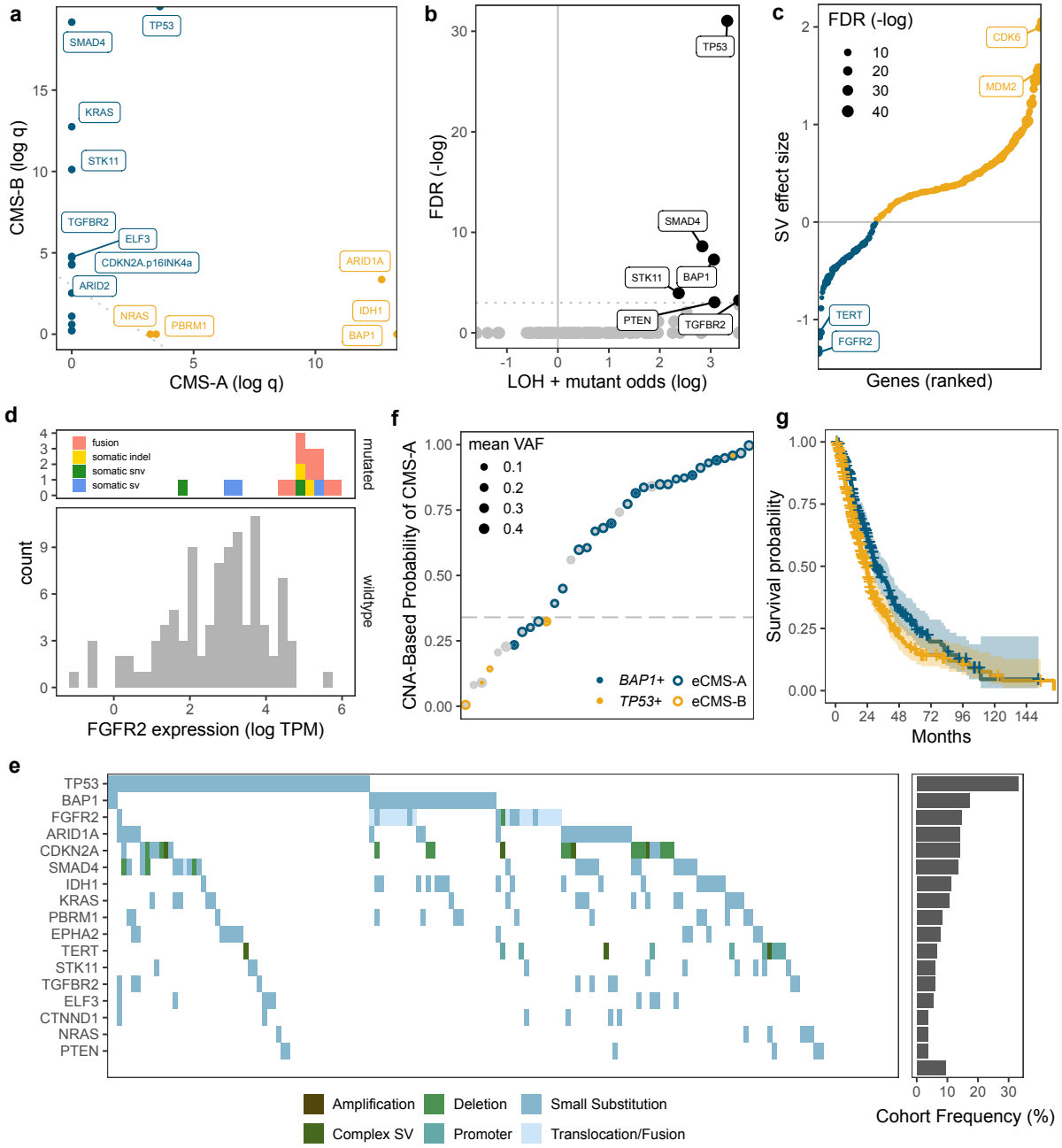
a
CMS-B (log q)
CMS-A (log q)
TP53
SMAD4
KRAS
STK11
TGFBR2
ELF3
CDKN2A.p16INK4a
ARID2
NRAS
PBRM1
ARID1A
IDH1
BAP1
b
FDR (-log)
LOH + mutant odds (log)
TP53
SMAD4
STK11
BAP1
PTEN
TGFBR2
c
FDR (-log)
10
20
30
40
SV effect size
Genes (ranked)
CDK6
MDM2
TERT
FGFR2
d
fusion
somatic indel
somatic snv
somatic sv
mutated
wildtype
count
FGFR2 expression (log TPM)
f
mean VAF
0.1
0.2
0.3
0.4
CNA-Based Probability of CMS-A
BAP1+
TP53+
eCMS-A
eCMS-B
g
Survival probability
Months
e
TP53
BAP1
FGFR2
ARID1A
CDKN2A
SMAD4
IDH1
KRAS
PBRM1
EPHA2
TERT
STK11
TGFBR2
ELF3
CTNND1
NRAS
PTEN
Amplification
Deletion
Small Substitution
Complex SV
Promoter
Translocation/Fusion
Cohort Frequency (%)

Figure S3

Fig. S3 | Extended data on drivers by CMS

a, genes showing significantly elevated dN/dS ratios in analysis within eCMS. **b,** over-representation of loss-of-heterozygosity in mutated cases for genes with significant mutational over-representation in previous analysis before multiple testing correction. **c,** structural variant expression-based quantitative trait loci (sv-eQTL) effective size for eCMS-A and -B. **d,** *FGFR2* expression in transcripts per million (TPM) for wildtype and mutated cases, across multiple mutation types (color). **e,** oncoplot of small substitutions and structural variants across genes with significant mutational over-presentation, by mutation type; cohort frequency in right panel . **f,** CNA-based assignment of gCMS-A (against gCMS-B) for samples from Farshidfar et al., by driver, eCMS (color) and mean variant allele frequency (size). **g,** Kaplan-Meir analysis of survival by gCMS in Song Y. *et al.*

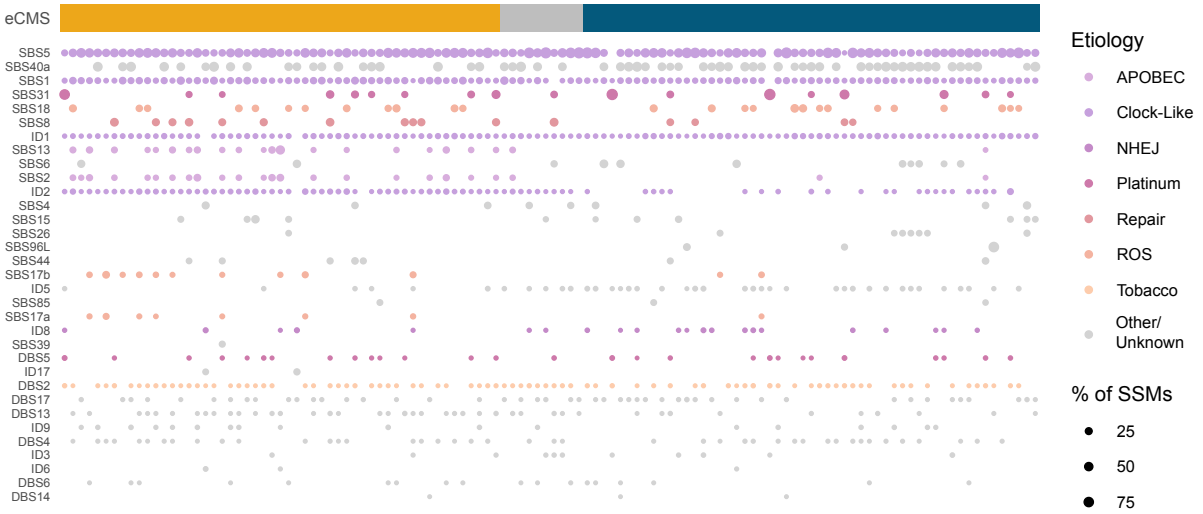
eCMS
SBS5
SBS40a
SBS1
SBS31
SBS18
SBS8
ID1
SBS13
SBS6
SBS2
ID2
SBS4
SBS15
SBS26
SBS96L
SBS44
SBS17b
ID5
SBS85
SBS17a
ID8
SBS39
DBS5
ID17
DBS2
DBS17
DBS13
ID9
DBS4
ID3
ID6
DBS6
DBS14
Etiology
APOBEC
Clock-Like
NHEJ
Platinum
Repair
ROS
Tobacco
Other/ Unknown
% of SSMs
25
50
75

Figure S4

Fig. S4 | Mutational signatures

full mutational signature assignments across samples and signatures, size as proportion of mutations in sample, clustered by mutational profile, top panel showing eCMS.

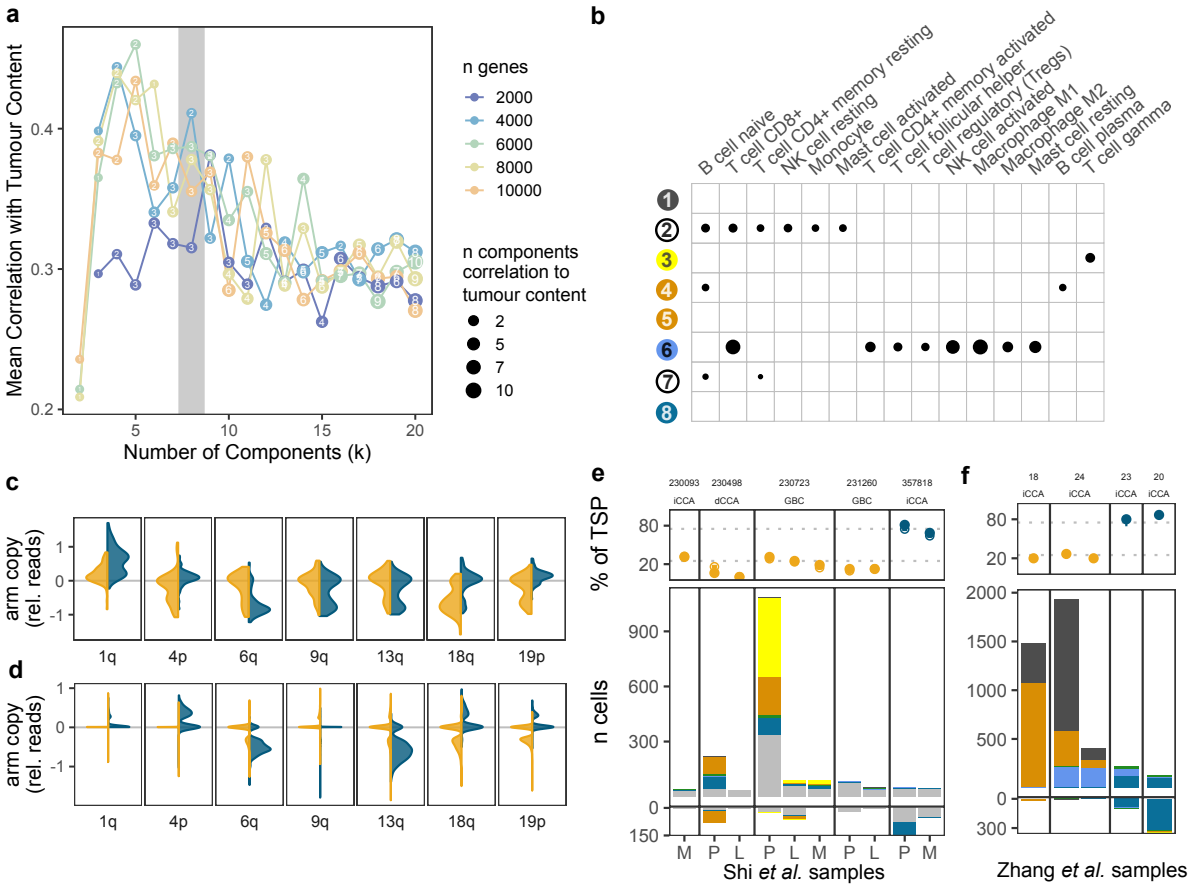
a
Mean Correlation with Tumour Content
Number of Components (k)
n genes
2000
4000
6000
8000
10000
n components correlation to tumour content
2
5
7
10
b
B cell naive
T cell CD8+
T cell CD4+ memory resting
NK cell resting
Monocyte
Mast cell activated
T cell CD4+ memory activated
T cell follicular helper
T cell regulatory (Tregs)
NK cell activated
Macrophage M1
Macrophage M2
Mast cell resting
B cell plasma
T cell gamma
c
arm copy (rel. reads)
1q
4p
6q
9q
13q
18q
19p
d
e
% of TSP
n cells
Shi et al. samples
f
Zhang et al. samples

Figure S5

Fig. S5 | Expression deconvolution and single cell RNAseq

a, parameters for manual optimization of gene count (color) and component count (k) for non-negative matrix factorization of counts per million (CPM) expression data, using mean correlation of all components with a correlation with WGS-estimate of sample tumour content greater than 0.2. Vertical gray bar shows chosen optimal k. **b,** correlation between signature weights from NMF and immune cell decomposition scores from *cibersort* by cell type. **c-d,** mean relative arm copy number for chromosome arms in the gCMS (CNA) model for **c,** cohort-level bulk WGS data, and **d,** single cells with expression of epithelial markers from Song G. *et al.* **e-f,** bootstrapped pseudobulk eCMS assignment with proportion of genes in TSP agreeing with call, and tallies of aneuploid cells by top NMF-expression signature assignment for scRNA of tumor samples from **e**, Shi *et al.*, and **f,** Zhang *et al.*

Supplementary Tables

Table S1. Cohort characteristics

		Overall	Surgical	Unresected
		n = 169	n = 97	n = 55
Age at Diagnosis		59.2 (12.5)	59.4 (12.9)	59.0 (11.8)
	Unknown	2	1	0
Sex assigned at birth				
	Female	71 (43%)	50 (52%)	20 (36%)
	Male	96 (57%)	46 (48%)	35 (64%)
	Unknown	2	1	0
Location				
	CCA NOS	1 (0.6%)	0 (0%)	1 (1.8%)
	Distal	9 (5.3%)	7 (7.2%)	2 (3.6%)
	Gallbladder	18 (11%)	9 (9.3%)	6 (11%)
	HCC-CCA	11 (6.5%)	6 (6.2%)	4 (7.3%)
	Intrahepatic	104 (62%)	62 (64%)	33 (60%)
	Perihilar	26 (15%)	13 (13%)	9 (16%)
Stage at Diagnosis				
	I & II	51 (44%)	44 (72%)	7 (14%)
	III	19 (17%)	10 (16%)	7 (14%)
	IV	45 (39%)	7 (11%)	35 (71%)
	Unknown	54	36	6
Nodes at Diagnosis				
	N0	73 (55%)	54 (69%)	16 (33%)

	N1	45 (34%)	13 (17%)	28 (58%)
	N2	2 (1.5%)	1 (1.3%)	1 (2.1%)
	NX	13 (9.8%)	10 (13%)	3 (6.3%)
	Unknown	36	19	7
Metastases at Diagnosis				
	M0	87 (66%)	67 (92%)	16 (31%)
	M1	44 (34%)	6 (8.2%)	35 (69%)
	Unknown	38	24	4

Table S2. BTC expression classifiers

Reference	Classes	Classifier data source	DOI
Andersen *et al.* 2012	2	Supplementary Table 1	10.1053/j.gastro.2011.12.005
Sia *et al.* 2013	2	Supplementary Table 2	10.1053/j.gastro.2013.01.001
Ahn *et al.* 2019	2	Supplementary Table 1	10.1007/s12072-019-09954-3
Montal *et al.* 2020	4	Table S11	10.1016/j.jhep.2020.03.008
Nepal *et al.* 2021	3	Table S6	10.1016/j.jhep.2020.11.033
Song G. *et al.* 2022	2	Supplementary Data 5	10.1038/s41467-022-29164-0
O'Rourke *et al.* 2024	2	Table S2	10.1136/gutjnl-2023-330748
Fan *et al.* 2024	2	Supplementary Figure 7	10.1038/s41467-024-44748-8
Vijay *et al.* 2025	4	Table S5 | RNA_cluster_signature	10.1101/2024.07.04.601970 (bioRxiv)

Table S3. Small substitution drivers

gene	chr.	n	Element (FDR)	dN/dS (q)	LOH (FDR)	LOH n
TP53	17	59	5.10E-72	0	2.50E-15	52
BAP1	3	30	1.60E-25	0	5.10E-05	28
ARID1A	1	28	4.70E-07	2.10E-12	0.33	20
SMAD4	18	22	1.10E-13	0	1.40E-05	18
IDH1	2	19	6.00E-15	0	0.78	0
KRAS	12	18	6.20E-22	0	1	2
EPHA2	1	14	2.80E-06	0.00026	0.57	10
PBRM1	3	14	0.00012	2.10E-06	0.0046	14
STK11	19	12	1.30E-06	2.10E-06	0.0014	8
ELF3	1	11	0.0028	1.70E-07	1	1
TGFBR2	3	9	0.31	0.0017	0.0029	11
CTNND1	11	7	1	0.044	1	0
NRAS	1	6	0.00079	0.031	0.53	0
PTEN	10	6	0.014	0.12	0.0036	5