Integrated whole-genome and transcriptome sequencing reveals divergent evolutionary processes across biliary tract cancer subtypes

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], Jingxiong 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

Conflicts of interest

GO: Consulting and advisory role: AstraZeneca, Roche, Servier, Incyte, MSD. Grants: Roche, AstraZeneca. Travel: MSD, Takeda, Roche; **TJP** has provided consultation for AstraZeneca, Chrysalis Biomedical Advisors, Guidepoint Global, Roche, and Merck (compensated); and receives research support (institutional) from AstraZeneca and Roche/Genentech; **GS**: Grant Support from Bayer, Roche. Consultant for Novartis, Integra, Roche, AstraZeneca, Chiesi, Eisai, Natera, HepaRegeniX. Research Collaborations with AstraZeneca, Natera, Roche-Genentech. **AV**: Consultancy and advisory role: Roche, AstraZeneca, Bohringer- Ingelheim, Ipsen, Incyte, Cogent, EISAI, Zymeworks, Biologix, BMS, Terumo, Elevar, Servier, MSD, Tahio, Jazzpharma,

Medivir, Abbvie, Tyra, Falk, Janssen, Lilly; **RG**: Received a graduate scholarship from Pfizer and provided consulting or advisory roles for Astrazeneca, Tempus, Eisai, Incyte, Knight Therapeutics, Guardant Health, and Ipsen. **RJ**: Advisory role: Merck. Research funding: Merck, Astra Zeneca, Camurus. **EE:** Consultancy role: BMS, Zymeworks, Adaptimmune, Astellas, Beigene, Viracta Tx, Novartis, Jazz, Natera, Abbvie, Daichi-Sankyo, Roche, Amgen. Research funding: BMS, Zymeworks, Amgen, Astra Zeneca, Jazz, Bold Therapeutics, Arcus Biosciences. Steering: Astra Zeneca, Jazz.

Abstract

Biliary tract cancer (BTC) comprises a family of rare malignancies subclassified by anatomy and pathology. However, this scheme may obscure shared biology and limit patient stratification. We therefore performed whole-genome and transcriptome sequencing of 169 tumors enriched for tumor cells by laser capture microdissection to identify shared programs in BTC. Network integration across transcriptomic classes identified two consensus cancer subtypes (CCS). CCS segregates with anatomical location of primary tumor and expressed gene marker analyses suggest subtypes reflect tumor cell of origin differences. CCS display strikingly divergent molecular landscapes, explaining more variance than anatomical location of primary tumor. CCS-B tumors are mutationally loaded with clock-like and APOBEC signatures and extrachromosomal DNA, whereas CCS-A tumors are characterised by chromosome-arm deletions and higher levels of subclonality. We show harnessing the genomic and transcriptomic diversity of BTC uncovers novel biology and improves stratification.

Significance

We provide evidence that biliary tract consensus cancer subtypes define fundamentally different cancers, with diverging modes of evolution stemming from distinct cells of origin

Introduction

Biliary tract cancer (BTC) is a rare and highly fatal disease [1,2]. Progress in understanding treatment response and identifying actionable biomarkers has been hindered by its extensive clinical and molecular heterogeneity [3,4]. The different staging across BTCs showcases its heterogeneity, with distinct systems for intrahepatic, perihilar, distal, and gallbladder cancers. The multitude of anatomical, molecular, and etiological factors makes it difficult to isolate the impact of any single variable, limiting the interpretability of traditional comparative analyses.

Emerging evidence suggests that many of these heterogeneous features covary, raising the possibility that they reflect deeper, latent organizational principles within BTC biology. Certain driver mutations tend to co-occur more than others, such as *FGFR2* fusions with *IDH1* mutations or *TP53* with *KRAS* mutations [5]. The former tend towards intrahepatic tumors while the later are more common in extrahepatic tumors. Patients with primary sclerosing cholangitis (PSC) or liver fluke infections, both of which are leading BTC risk factors, tend to have *KRAS* or *TP53* mutant tumors [6,7]. Several gene expression classifiers with prognostic implications show different rates of driver mutations [8–12], but how they relate to each other remains underexplored.

To date, clinical BTC classification has mostly been anchored to anatomic origin or mutations in specific genes [13]. In this study, we tested whether BTC heterogeneity can be explained by coherent latent axes and evaluate their potential to unify diverse clinical and genomic factors under a tractable biological framework. To this end, we performed whole genome and transcriptome sequencing on 169 BTC tumor samples annotated with clinical information and characterized their expression profiles, genetic drivers, and mutational signatures (Fig. 1a).

Results

Cohort description

We analyzed the whole genome from one tumor specimen for each of 169 patients with BTC, including 72 previously reported cases [13,14] and 97 new cases recruited prospectively. Tumors also underwent whole transcriptome sequencing (WTS), when enough tissue remained (n = 134). 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. Primary tumors spanned the biliary tree, including 104 intrahepatic, 11 combined hepatocellular-cholangiocarcinoma, 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 tumor specimens (n = 105, 62%) were treatment-naive (Fig. S1a). Fresh frozen samples (n = 148) were preferred over formalin-fixed paraffin-embedded tissue (n = 21; Fig. S1b). As BTC typically exhibits desmoplasia and low tumor cellularity, most samples were enriched using laser-capture microdissection (n = 153).

Network integration of published transcriptomic classifiers reveals two subtypes

Expression-based classifiers of BTC have been previously proposed but studied in isolation. To explore overlap between previous classifiers, we classified all 116 fresh frozen samples with WTS in our cohort into classes from nine independent classifiers (Table S2; Fig. S2a). Next, we constructed a similarity network graph in which nodes represented classes and edge weights reflected the frequency at which samples were assigned to both classes connected by that edge (Fig. S2b). Edge weights between classes from the same classifier are zero by definition. After network optimization (see *Methods*, Fig. S2c), we performed Markov clustering on the class network [15]. Ten classes from five classifiers resolved into two significantly interconnected clusters of classes (p = 0.002, hypergeometric test, Fig. 2a) suggesting high overlap between existing classifiers. We provisionally named the clusters consensus cancer subtype-A and -B (CCS-A and -B).

Two classifiers [8,9] in the final network were trained to maximize survival differences among patients. To confirm CCS retained this ability, we first developed a top scoring pair model[16] for single-sample expression-based CCS assignment (eCCS). We found relative expression at 16 pairs of genes could effectively distinguish samples according to CCS (Table S3; Fig. S2d, e). Testing the eCCS model in a held-out test set achieved 100% accuracy (95% C.I.: [82 - 100%]). Returning to survival differences, we found eCCS-A was associated with significantly longer overall survival with first-line palliative systemic therapy than eCCS-B (p = 0.007; median survival of 22.4 and 15.0 months respectively; Fig 2b). Assigning eCCS to tumor samples from two external cohorts, patients with CCS-A tumors had significantly longer survival after surgery in the 147 intrahepatic cholangiocarcinoma patients from Dong *et al.* [17] (p = 9.5e-05) and from diagnosis for 62 BTC patients from Jusakul *et al.* [18] (p = 0.002) (Fig S2f-g).

To extend analysis to datasets without gene expression data, we also trained a genetics-based model (gCCS) to predict CCS based on median relative arm-level copy number (*i.e.* log tumor / normal read counts; Table S4). On a held-out test dataset (n = 22), a model predicting gCCS had an area under the receiver operator curve of 0.87 (Fig 2c). Validating the gCCS model on the data of Farshidfar *et al.* [19], the area under the curve was 0.95 (sensitivity = 1 where specificity = 0.85). Stratifying patients from Song Y. *et al.*[20] by gCCS using their genetic panel data (n = 534), we saw significantly longer overall survival in gCCS-A (p = 0.0029; median survival of 29.9 and 23.6 months; Fig S2h). These results suggest the eCCS classifier outperforms the gCCS model in accuracy, but both models capture survival differences.

Three classifiers [9–11] were specifically trained on intrahepatic cholangiocarcinomas. Exploring the alignment of anatomical location with CCS, we found eCCS-A samples appeared exclusively in intrahepatic cholangiocarcinomas or combined hepatocellular-cholangiocarcinoma, while eCCS-B spanned the biliary tract (Fig. 2d). The site of biopsy from metastases did not affect the association between eCCS and primary site of the tumor: all 17 liver metastases from extrahepatic primaries were assigned to eCCS-B. Analyzing 11 primary and metastatic paired patient samples, all pairs exhibited concordant eCCS assignment despite different biopsy sites. Within intrahepatic cholangiocarcinomas, the classifier from Song G. *et al.*[11] was trained to distinguish between small and large duct pathologies. Using the cohort of Jolissaint *et al.* [21], the gCCS model recovered an enrichment of large duct intrahepatic cholangiocarcinoma within gCCS-B (n = 12, p = 0.01), while the small duct histopathology was significantly enriched in gCCS-A (n = 117, p = 0.01; Fig. 2e).

Given the convergence of CCS onto both an anatomical location and a survival difference, we tested whether anatomical location or CCS better predicted survival. In a multivariable model including disease stage, eCCS was significantly associated with survival time after start of palliative treatment ($p = 0.017$) while anatomical location was not (Fig. 2f). Similarly, in the data from Song Y. *et al.*[20], gCCS was significantly associated with survival time after surgery ($p = 0.005$), while anatomical location was not (Fig. 2g).

Subtypes reflect different tumor intrinsic transcriptional programs

CCS groups 32% ($n = 23$) of intrahepatic cholangiocarcinomas with perihilar and extrahepatic cholangiocarcinomas, suggesting higher-order biological structure not captured by anatomy alone. We tested whether grouping tumors by CCS explained more variation in gene expression than anatomical location. In a redundancy analysis, primary location and eCCS both explained significant variance in the first ($p = 0.023$ and $p = 0.001$) and third (both $p < 0.001$) principal components of gene expression (Fig. 2a). However, on both axes, eCCS explained more variance in gene expression than anatomical location of primary tumor (Fig. 2b). These results show CCS captured more transcriptional variation than anatomical location.

Next, we enquired whether CCS reflects tumor intrinsic features or the tumor microenvironment. The top scoring pairs identifying eCCS-A include canonical cholangiocyte markers (*DCDC2*, *BICC1*, *ONECUT1*) suggestive of a signal from biliary epithelial cells. While genes from top scoring pairs for eCCS-B did not help resolve signal identity, differential gene expression showed eCCS-B tumors had higher expression of gastrointestinal epithelium markers (*CDX2*, *KRT19*, *KRT20*, *CD44*), which are also used to define intraductal papillary neoplasms of the bile ducts [22,23]. These marker enrichment profiles were replicated in methylation data from Jusakul *et al.*[18] and protein expression data from Dong *et al.*[17] (Fig 3b, c). The enrichment of epithelial markers across data modalities supports that CCS reflect different epithelial-derived tumor cell populations.

Because bulk data collapses expression profiles from multiple cell types, we performed virtual microdissection using non-negative matrix factorization (NMF) to separate expression signatures. Varying the expected number of signatures, eCCS remained a top differentiating factor (Fig. S3a). When optimizing for components that best correlated with tumor content, the data were best explained by eight signatures across 4,000 genes (see *Methods*; Fig. S3b). Gene set enrichment analysis found two NMF signatures (signatures 3 and 4) were significantly enriched for eCCS-B top scoring pair markers (normalized enrichment scores = 2.0 and 2.0, false discovery rate [FDR] = 3e-05 and 2e-05, respectively; Fig. 3e). Both signatures were also enriched for scRNA markers of common bile duct epithelium [24] (FDR = 2.7e-07 and 1.4e-09) and genes corresponding to goblet cell proteins from the Human Protein Atlas [25] 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 [26]. With the top scoring pair markers for eCCS-A, gene set enrichment analysis found one significantly enriched signature: signature 8 (normalized enrichment scores = 2.2, FDR = 2.7e-09). Gene set enrichment analysis of scRNA markers for signature 8 found enrichment of intrahepatic duct markers (FDR = 9e-10), and, of Human Protein Atlas markers, we found overrepresentation of genes with protein expression in renal proximal tubules (FDR = 3.7e-05).

NMF also uncovered two immune signatures. Weights for signature 2 were correlated with transcriptome-derived estimates of resting CD4+ memory cells, resting NK cells, monocytes, and activated mast cells (see *Methods*; Fig. S3c), suggestive of an immunologically cold environment. Signature 6 was associated with activated CD4+ memory cells, helper follicular T cells, regulatory T cells, activated NK cells, M1 macrophages and resting mast cells, suggestive of immune hot tumors. We saw the majority of variance across these two immune components in specimens with low tumor content (<=65%), while these signals were virtually absent from specimens with high tumor content (>65%; Fig. 3f). Mapping samples to the stroma-tumor-immune microenvironment (STIM) classification framework of Martin-Seranno *et al.*[27], we similarly saw a strong impact of tumor content on class assignment (Fig. 3g). In mid-high tumor content tumors, eCCS-A tumors map to the 'hepatic stem-like' class while eCCS-B tumors map to the 'tumor classical' or 'desert-like' classes, while in low tumor content specimens, eCCS-A tumors mapped to the 'immune classical' class and eCCS-B tumors were assigned to 'inflammatory stroma'. Regardless of tumor content, we did not see associations between CCS and immune signatures derived from NMF, suggesting the immune environment is independent of CCS.

We next used single cell RNAseq from Song G. *et al.* [11] to confirm that CCS signal originated from individual tumor cells. We first inferred aneuploidy for each cell, a marker of malignancy [28], and used the inferred copy number profiles to predict gCCS for individual cells. Because gene dropout made eCCS assignment by top scoring pairs unreliable at single cell resolution, we projected the richer NMF-derived gene expression signatures onto single cells and reserved eCCS assignment for pseudobulked tumors. Tallying cells by their top signature [29], we found most gCCS-B cells showed high levels of signature 4, and derived from pseudobulked eCCS-B tumors, while gCCS-A cells upregulated signature 8 and appeared in pseudobulked eCCS-A tumors (Fig 3h). Quantifiably similar conclusions could be made with the scRNA BTC datasets from Shi *et al.* [30] and Zhang *et al.* [31] (Fig S3d, e). These results suggest CCS captures tumor cell-intrinsic transcriptional states.

Subtypes harbor distinct genomic drivers

Studies of co-alteration patterns of mutations in pan-cancer driver genes have recurrently found two genetic profiles in BTC [5] but previous studies have been underpowered to define rare BTC-specific drivers[32,33]. To test whether the CCS framework aligns with previously identified genetic profiles and to cement these mutations as statistically supported drivers, we turned to our whole genome sequencing for all 169 samples, calling BTC drivers *de novo*. 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 FDR in one approach (Fig. 4a, Table S5). 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 substitution driver analysis restricted within eCCS-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 eCCS-A (Fig. S4a).

Next, we explored genes impacted by recurrent structural variants. To enrich for functional impact, we tested structural variant breakpoints overlapping expression quantitative trait loci [34]. Top structural variant expression quantitative trait loci within eCCS-A were at *FGFR2* (n = 19), and *TERT* (n = 3), while top loci for eCCS-B were at *CDK6* (n = 8) and *MDM2* (n = 5) (Fig. S4b). Amplicon reconstructions from copy number profiles found the regions encompassing *MDM2* and *CDK6* were commonly extrachromosomal DNA (n = 4 of 5 and n = 3 of 8, respectively). Because most high effect size structural variants were rare (<5% of cases), we formally assessed recurrence with a Q-Q enrichment analysis [35]. *FGFR2* was the only gene with somatic structural variants showing significant expression quantitative trait association and statistical recurrence (FDR = 2.4e-04).

We explored how mutations in drivers were co-distributed among tumors and how this related to CCS (Fig. S4c). Testing how often samples were mutated versus wildtype, we recovered a nexus of mutual exclusively defined by *TP53*, *SMAD4*, and *KRAS* mutations in eCCS-B cases, in contrast to mutations in *IDH1*, *FGFR2*, and *BAP1* in eCCS-A (Fig. 4b, c). We also found novel associations such as *NRAS* with eCCS-A and *STK11* with eCCS-B.

Next, we examined copy number variants across BTC. Under Knudson's two-hit hypothesis [36], copy-loss events leading to loss-of-heterozygosity should converge on tumor suppressor genes. Six driver genes showed significantly more loss-of-heterozygosity events in mutated compared to wildtype cases, with loss-of-heterozygosity occurring at the highest rates with mutations in *TP53* (52 loss-of-heterozygosity events among 59 mutations, FDR = 2.5e-15; Fig. S4d). 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.

Evaluating recurrence of copy number alterations [37], we found well-characterized oncogenes such as *ERBB2* and *MYC* as well as recurrent homozygous deletions of tumor suppressor gene *CDKN2A* (Fig. 4d). GISTIC [37] within CCS revealed divergent copy number profiles: recurrent narrow amplifications were over-represented in eCCS-B, including *GATA6*, *CCND1* and *CCNE1*. eCCS-B tumors were also more likely to exhibit extrachromosomal DNA (p = 0.0003; Fig. S4e). In contrast, eCCS-A commonly showed chromosome-arm scale heterozygous deletions (single arm losses). Tumor suppressor genes in other cancer types with changes in gene expression in eCCS-A included *ERFFI1* and *IRF2*, but we also note recurrent deletion of 6q and amplification of 1p21 without conclusively identifiable drivers.

The segregation patterns of drivers may play a role in differences in survival outcomes between CCS. Because *TP53* is a prognostic factor in BTC [38], we tested whether eCCS was still prognostic in *TP53* wildtype cases. Subsetting our palliative cohort to *TP53* wildtype cases, we saw a trend towards longer survival in eCCS-A, but the difference was not significant (p = 0.27), potentially because only 14 eCCS-B cases remained. In the 71 *TP53* wildtype intrahepatic cholangiocarcinomas from Dong *et al.*[17], eCCS was significantly associated with overall survival after surgery (Fig. S4f). Furthermore, in these data, a multivariable Cox proportional hazards model including eCCS, *TP53* status and tumor stage found a significant association with survival and eCCS (p = 0.00025) but not *TP53* (p = 0.56). Another possible factor confounding survival with CCS was the strong segregation of clinically targetable alterations in CCS-A, as both *FGFR2*-fusion and *IDH1*-mutated tumors

have FDA-approved therapies. In our cohort, two cases were treated with pemigatinib (an FGFR2-inhibitor), while three were treated with an IDH1-inhibitor on trial in the palliative setting. When these cases were removed from the analysis, we still observed a significant difference in OS by eCCS (p = 0.017) while anatomical location remained not significant.

Divergent evolutionary dynamics between subtypes

Turning to the broader mutational context, we observed a 1.6-fold increase in tumor mutational burden (TMB) in eCCS-B over A (p = 3.3e-06, Fig 5a). These differences remained significant after excluding highly mutated samples (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 CCS-B over CCS-A (Fig. S4g). 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). eCCS was a better predictor of TMB than primary location: comparing results of independent linear models, eCCS explained 15.8% of the variance while primary location explained 6.3%, while in a two-factor ANOVA eCCS was significant (p = 0.0017) while primary location was not (p = 0.9).

We next compared whether the increased burden could be explained by specific mutational signatures. *De novo* mutational signature extraction identified 19 SBS, seven double basepair substitution (DBS) and eight ID signatures (Fig. S5). Comparing signature contributions between CCS using a linear model, eCCS-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 eCCS-A tumors showed enrichment of ID5 (unknown etiology; p = 2.9e-05; Fig 5b). Further, CCS-associated 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 eCCS-A from eCCS-B tumors (p = 1.3e-09; Fig. 5c). 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). Once more, CCS better explained variance than primary location: permutation analysis across PC1 and PC2 found CCS explained 12.5% of the variance in mutational signatures (p = 0.001) while primary location explained 6.8% (p = 0.14). These results suggest not only significant differences in mutational profiles between CCS, but that these differences were 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, eCCS-A showed a significantly higher number of subclones (p = 0.03) and Shannon diversity index (p = 0.008) than eCCS-B. Breaking down the distribution of cancer cell fraction, a larger proportion of variants in eCCS-A belonged to rare clones (< 0.5 cancer cell fraction), while the majority of variants in eCCS-B were clonal or nearly clonal (Fig. 5d). Among clonal mutations, signatures showing CCS-bias occurred early in that subtype's evolution: relatively more APOBEC and clock-like signatures were earlier than later in eCCS-B (p = 0.00024 and p = 6.8e-09, respectively) with a similar finding for ID5 in eCCS-A (p = 2.6e-05; Fig. 5e). These results suggest these signatures may be important in early stages of subtype evolution.

Discussion

Molecular subtyping has stratified many cancers into biologically and clinically meaningful groups [39–41]. However, in BTC, classification has been anchored to anatomy or mutations in specific genes. In this study, we integrated published transcriptomic literature to define two transcriptomic subtypes, CCS-A and CCS-B. By integrating whole-genome and transcriptome sequencing in a large cohort, we demonstrate that the CCS capture more variation across gene expression, driver mutations, mutational signatures, and clinical outcomes than anatomical classification, supporting that the CCS represent a high-order biological stratification. We demonstrated that the CCS subtypes reflect different tumor-intrinsic epithelial expression programs. The association between CCS-A and small duct single-cell markers and CCS -B with extrahepatic markers, combined with the distinct embryological origin differences between these tissues [42], suggest that the CCS may stratify BTC based on distinct cell of origin.

Whole genome sequencing supports divergent evolutionary trajectories between CCS. We identified drivers in CCS-B newly associated with BTC, including *EPHA2.* We also revealed putative tumor suppressors in CCS-A, including *ERRFI1*, which has potential therapeutic implications [43] and *IRF2* which emerged as a candidate gene for BTC predisposition in a germline genome-wide association study [44]. We also found distinct copy number profiles between subtypes, with CCS-B enriched for narrow high-copy number changes. Extrachromosomal DNA provides a mechanism for oncogene activation in CCS-B, exemplified by amplification and overexpression of *MDM2*. CCS-B tumors have higher mutational burdens, mostly driven by clock-like or APOBEC-associated signatures, starting early in tumor evolution. The association between the enrichment of molecular clock signatures and precursor lesions and metaplasia in CCS-B could provide fruitful insights in the early stages of evolution of this subtype. CCS-A tumor cells exhibited large-scale copy number alterations with few well-characterized drivers. We also see an enrichment of subclones and a lower mutational burden; whether CCS-A tumors have a lower mutation rate or exhibit more effective selection against new mutations will require further study.

The CCS framework provides an organizing system for BTC that balances the trade-off between grouping similar tumors to maintain statistical power in clinical trials for these rare cancers [45] and distinguishing features of meaningful subtypes [46]. We anticipate that larger BTC cohorts will reveal further substratification within the subtypes. We also expect stratification exists within the tumor microenvironment as has been shown by immunohistochemistry [47].

In conclusion, our data provide the first comprehensive integration of multi-omic and clinical evidence to define consensus cancer subtypes for BTC. 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 this historically heterogeneous disease.

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. Sample and basic data processing methods can be found in the supplementary materials.

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.

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. Count data were normalized using *Deseq2* v 1.48.1 [48] and scaled using *ematAdjust* from *CMScaller* v0.99.2 [49] . 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 [15].

To train a top-scoring pair algorithm, we restricted the training set to samples that were assigned to the same CCS across every classifier. The algorithm was trained on 80% of the data (n = 70) with the *SWAP.KTSP.Train* function of *switchBox* v 1.44.0 [16] 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 CCS 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).

To train a copy number-based CCS 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 stepwise 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 gCCS classifier, data from Farshidfar *et al.* [19] and Song Y. *et al.* [20] were downloaded from cbioportal.org . Methods for processing of external datasets can be found in supplementary materials.

Gene expression profiles and virtual microdissection

Differential gene expression between samples from each CCS 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[50] 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 maximized. 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.* [24] and single nucleus data from Table S10 from Andrews *et al.* [51] and run on *GSEA* from *clusterProfiler* v 4.16.0 [52]. Human Protein Archive enrichment was performed using *gost* from *gprofiler2* v 0.2.3. *CIBERSORT* correlations were obtained from *immunedeconv* v 2.1.0 [53] deconvolution with "cibersort_abs".

Genomic variant calling

We used *ActiveDriverWGS* v 1.2.1 [54] 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 [55]. Promoter regions were tested exclusively with *ActiveDriverWGS*. All significantly recurrent hits were manually reviewed in integratedGenomeViewer [56] to exclude germline variants.

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 [35]. Driver co-occurrence significance was corrected for variant frequency using *Rediscover* v 0.3.2 [57].

We performed a Fisher's exact test querying whether any genes were enriched for loss-of-heterozygosity when also hit by a small substitution. *GISTIC* v2.0 [37] 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.

Mutational landscape

Mutational signature analysis was performed *de novo* using SigProfilerExtractor v 1.2.2 [58], 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. eCCS signature enrichment was performed using a binomial linear regression model [59], 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 [60]. Mutations were timed using the mutationTime function with n.boot = 10 and rho = 0 in *MutationTimer* v 1.0.2 [61]. Mutations where both clonality callers disagreed were discarded from clonality analyses. Clusters of putatively subclonal mutations at clonal fractions between 0.45-0.55 were removed as likely fixed heterozygous variants. Signature timing was tested by comparing between clonal early and clonal late the distributions in relative exposures in a Wilcoxon rank-sum test.

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 and caregivers who contributed to this research.

References

1. Vogel, A. *et al.* Biliary tract cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. *Ann. Oncol.* **34**, 127–140 (2023).
2. Banales, J. M. *et al.* Cholangiocarcinoma 2026: status quo, unmet needs and priorities. *Nat. Rev. Gastroenterol. Hepatol.* **23**, 65–96 (2026).
3. Ney, A. *et al.* Biliary Strictures and Cholangiocarcinoma – Untangling a Diagnostic Conundrum. *Front. Oncol.* **11**, 699401 (2021).
4. Macias, R. I. R. *et al.* The search for novel diagnostic and prognostic biomarkers in cholangiocarcinoma. *Biochim. Biophys. Acta BBA - Mol. Basis Dis.* **1864**, 1468–1477 (2018).
5. Kendre, G. *et al.* Charting co-mutation patterns associated with actionable drivers in intrahepatic cholangiocarcinoma. *J. Hepatol.* **78**, 614–626 (2023).
6. Chan-on, W. *et al.* Exome sequencing identifies distinct mutational patterns in liver fluke–related and non-infection-related bile duct cancers. *Nat. Genet.* **45**, 1474–1478 (2013).
7. Wang, X. *et al.* Clinicogenomic Characterization of Primary Sclerosing Cholangitis–Associated Biliary Tract Cancers. *Clin. Cancer Res.* **32**, 363–374 (2026).
8. Andersen, J. B. *et al.* Genomic and Genetic Characterization of Cholangiocarcinoma Identifies Therapeutic Targets for Tyrosine Kinase Inhibitors. *Gastroenterology* **142**, 1021-1031.e15 (2012).
9. 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.
10. Ahn, K. S. *et al.* Prognostic subclass of intrahepatic cholangiocarcinoma by integrative molecular–clinical analysis and potential targeted approach. *Hepatol. Int.* **13**, 490–500 (2019).
11. Song, G. *et al.* Single-cell transcriptomic analysis suggests two molecularly distinct subtypes of intrahepatic

cholangiocarcinoma. *Nat. Commun.* **13**, 1642 (2022).

12. 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.
13. 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).
14. 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).
15. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21**, 1350–1356 (2015).
16. 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).
17. Dong, L. *et al.* Proteogenomic characterization identifies clinically relevant subgroups of intrahepatic cholangiocarcinoma. *Cancer Cell* **40**, 70-87.e15 (2022).
18. Jusakul, A. *et al.* Whole-Genome and Epigenomic Landscapes of Etiologically Distinct Subtypes of Cholangiocarcinoma. *Cancer Discov.* **7**, 1116–1135 (2017).
19. Farshidfar, F. *et al.* Integrative Genomic Analysis of Cholangiocarcinoma Identifies Distinct IDH-Mutant Molecular Profiles. *Cell Rep.* **18**, 2780–2794 (2017).
20. Song, Y. *et al.* A Novel Approach to Quantify Heterogeneity of Intrahepatic Cholangiocarcinoma: The Hidden-Genome Classifier. *Clin. Cancer Res.* **30**, 3499–3511 (2024).
21. 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).

22. 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).

23. Zen, Y. & Akita, M. Neoplastic Progression in Intraductal Papillary Neoplasm of the Bile Duct. *Arch. Pathol. Lab. Med.* **148**, 989–996 (2024).

24. Sampaziotis, F. *et al.* Cholangiocyte organoids can repair bile ducts after transplantation in the human liver. *Science* **371**, 839–846 (2021).

25. Thul, P. J. & Lindskog, C. The human protein atlas: A spatial map of the human proteome. *Protein Sci.* **27**, 233–244 (2018).

26. Sakamoto, H. *et al.* A close relationship between intestinal metaplasia and Cdx2 expression in human gallbladders with cholelithiasis. *Hum. Pathol.* **38**, 66–71 (2007).

27. Martin-Serrano, M. A. *et al.* Novel microenvironment-based classification of intrahepatic cholangiocarcinoma with therapeutic implications. *Gut* **72**, 736–748 (2023).

28. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).

29. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).

30. Shi, X. *et al.* Single-cell atlas of diverse immune populations in the advanced biliary tract cancer microenvironment. *Npj Precis. Oncol.* **6**, 58 (2022).

31. Zhang, M. *et al.* Single-cell transcriptomic architecture and intercellular crosstalk of human intrahepatic cholangiocarcinoma. *J. Hepatol.* **73**, 1118–1130 (2020).

32. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385.e18 (2018).

33. 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).

34. Rheinbay, E. *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, 102–111 (2020).

35. Imielinski, M., Guo, G. & Meyerson, M. Insertions and Deletions Target Lineage-Defining Genes in Human Cancers. *Cell* **168**, 460-472.e14 (2017).

36. Knudson, A. G. Mutation and Cancer: Statistical Study of Retinoblastoma. *Proc. Natl. Acad. Sci.* **68**, 820–823 (1971).

37. 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).

38. Chae, H. *et al.* Therapeutic relevance of targeted sequencing in management of patients with advanced biliary tract cancer: DNA damage repair gene mutations as a predictive biomarker. *Eur. J. Cancer* **120**, 31–39 (2019).

39. Ceccarelli, M. *et al.* Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell* **164**, 550–563 (2016).

40. Notta, F. *et al.* A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns. *Nature* **538**, 378–382 (2016).

41. The Cancer Genome Atlas Research Network. Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. *N. Engl. J. Med.* **374**, 135–145 (2016).

42. 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).

43. 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).

44. Han, Y. *et al.* Uncovering the genetic landscape of cholangiocarcinoma and its subtypes via GWAS and integrative analyses. *Hepatology* https://doi.org/10.1097/HEP.0000000000001699 (2026) doi:10.1097/HEP.0000000000001699.

45. 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).

46. Chakravarty, D. & Solit, D. B. Clinical cancer genomic profiling. *Nat. Rev. Genet.* **22**, 483–501 (2021).

47. Job, S. *et al.* Identification of Four Immune Subtypes Characterized by Distinct Composition and Functions of Tumor Microenvironment in Intrahepatic Cholangiocarcinoma. *Hepatology* **72**, 965–981 (2020).

48. 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).

49. 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).

50. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).

51. 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).

52. 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).

53. 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).

54. 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).

55. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029-1041.e21 (2017).

56. 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).

57. Ferrer-Bonsoms, J. A., Jareno, L. & Rubio, A. Rediscover: an R package to identify mutually exclusive mutations. *Bioinformatics* **38**, 844–845 (2022).

58. Islam, S. M. A. *et al.* Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genomics* **2**, 100179 (2022).

59. Díaz-Gay, M. *et al.* Geographic and age variations in mutational processes in colorectal cancer. *Nature* **643**, 230–240 (2025).

60. Gillis, S. & Roth, A. PyClone-VI: scalable inference of clonal population structures using whole genome data. *BMC Bioinformatics* **21**, 571 (2020).

61. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).

Figures

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. CCS, consensus molecular subtype. HCC-CCA, combined hepatocellular-cholangiocarcinoma. eCCS: expression-based consensus molecular subtypes; gCCS: genetics-based consensus molecule subtypes.

Fig. 2 | Consensus cancer subtypes of biliary tract cancer

a, assignment similarity network between the CCS. **b**, overall survival after start of first-line palliative treatment by expression CCS in the present cohort. **c**, Accuracy of copy number alteration generalized linear model at predicting expression CCS (eCCS) and small mutation drivers for a test cohort (n = 24). **d**, counts of tumors classified by expression consensus subtype (CCS) into anatomical location of primary tumor on the biliary tract, as well as specimen location. **e**, Likelihoods for CCS-A associated across anatomical location of primary tumor for n = 129 samples from Jolissaint *et al.* (2021). **f-g**, Cox proportional hazards across **f,** expression CCS, primary tumor location, and tumor stage and, **g,** for genomic CCS (gCCS), location and grade on overall survival from diagnosis for unresectable disease for 534 samples from Song Y. *et al* 2024. eCCS, expression consensus molecular subtype. CCA: cholangiocarcinoma; combined hepatocellular-cholangiocarcinoma (HCC-CCA), iCCA: intrahepatic CCA; pCCA: perihilar CCA, dCCA: distal CCA, GBC: gallbladder cancer. TPM, transcripts per million. TSP, top scoring pairs. CCS, consensus cancer subtype.

Fig. 3 | *In silico* dissection of tissue and cell types for consensus cancer subtypes.

a-b, comparing principal components of variance in gene expression across tumors as **a,** principal components one and three (with inner color as primary location and outer color as eCCS) and **b,** r^2 of constrained variance in a redundancy analysis for location (□) and eCCS (□) across PC1 and 3. **c-d**, Differential **c,** methylation and **d,** gene expression between subtypes using the datasets from Jusakul *et al.* (2017) and Dong et *al.* (2021). **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-g,** Variance in tumor immune microenvironment across specimen tumor content using **f,** signatures 2 and 6 and **g,** the Stroma-Tumor-Immune Microenvironment classification of Martin-Serrano *et al.* (2023). **h**, eCCS per tumor by pseudobulk TSP, and tallies of cells by top NMF component split by single cell genetic CCS (gCCS). TSP: top scoring pair.

Fig. 4 | 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-occurrence in blue and mutual exclusion in red, and

significant associations circled in black. **c**, Log of the odds ratio for driver mutations by consensus subtype (CCS). **d**, Significance of copy number alteration recurrence analysis using GISTIC, across chromosomes by CCS and using all samples (globally). FDR: False discovery rate.

Fig. 5 | mutation profiles of biliary tract cancers.

a, Tumor mutational burden difference between consensus cancer subtypes (CCS). **b**, Association between CCS and mutational signatures for single basepair substitutions (●) and in/dels (○). **c** principal component analysis of mutational signatures showing platinum therapy exposed samples along the first principal component and CCS on the second. **d**, average number of mutations per subclone at each cancer cell fraction for each CCS. **e**, relative evolutionary timing of clonal mutations by mutational signature and CCS. SBS: single basepair substitutions; PC: principal component.

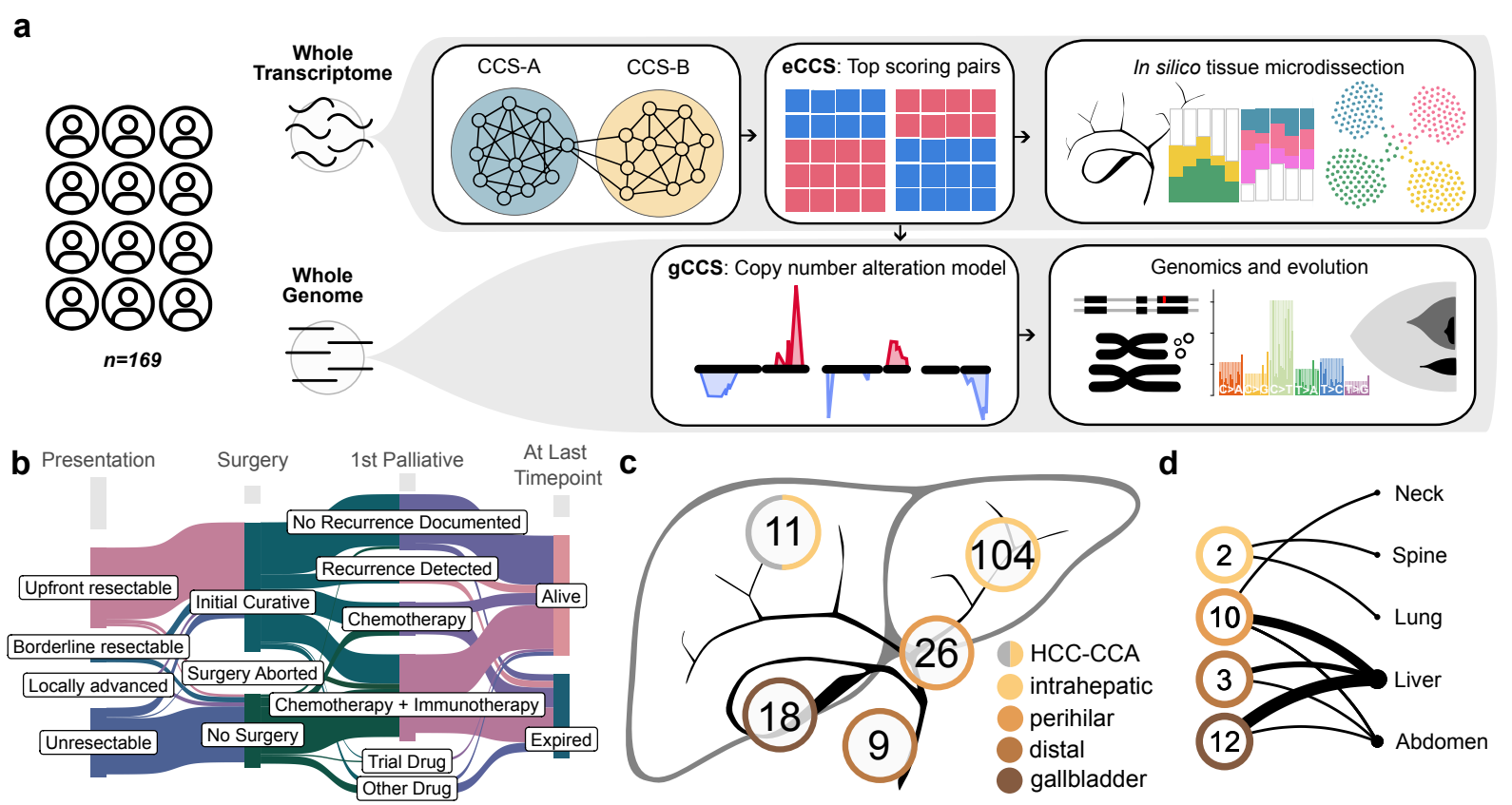
a
Whole Transcriptome
CCS-A
CCS-B
eCCS: Top scoring pairs
In silico tissue microdissection
n=169
Whole Genome
gCCS: Copy number alteration model
Genomics and evolution
b
Presentation
Surgery
1st Palliative
At Last Timepoint
No Recurrence Documented
Recurrence Detected
Upfront resectable
Initial Curative
Alive
Chemotherapy
Borderline resectable
Locally advanced
Surgery Aborted
Chemotherapy + Immunotherapy
Unresectable
No Surgery
Expired
Trial Drug
Other Drug
c
11
104
26
18
9
HCC-CCA
intrahepatic
perihilar
distal
gallbladder
d
2
10
3
12
Neck
Spine
Lung
Liver
Abdomen

Figure 1

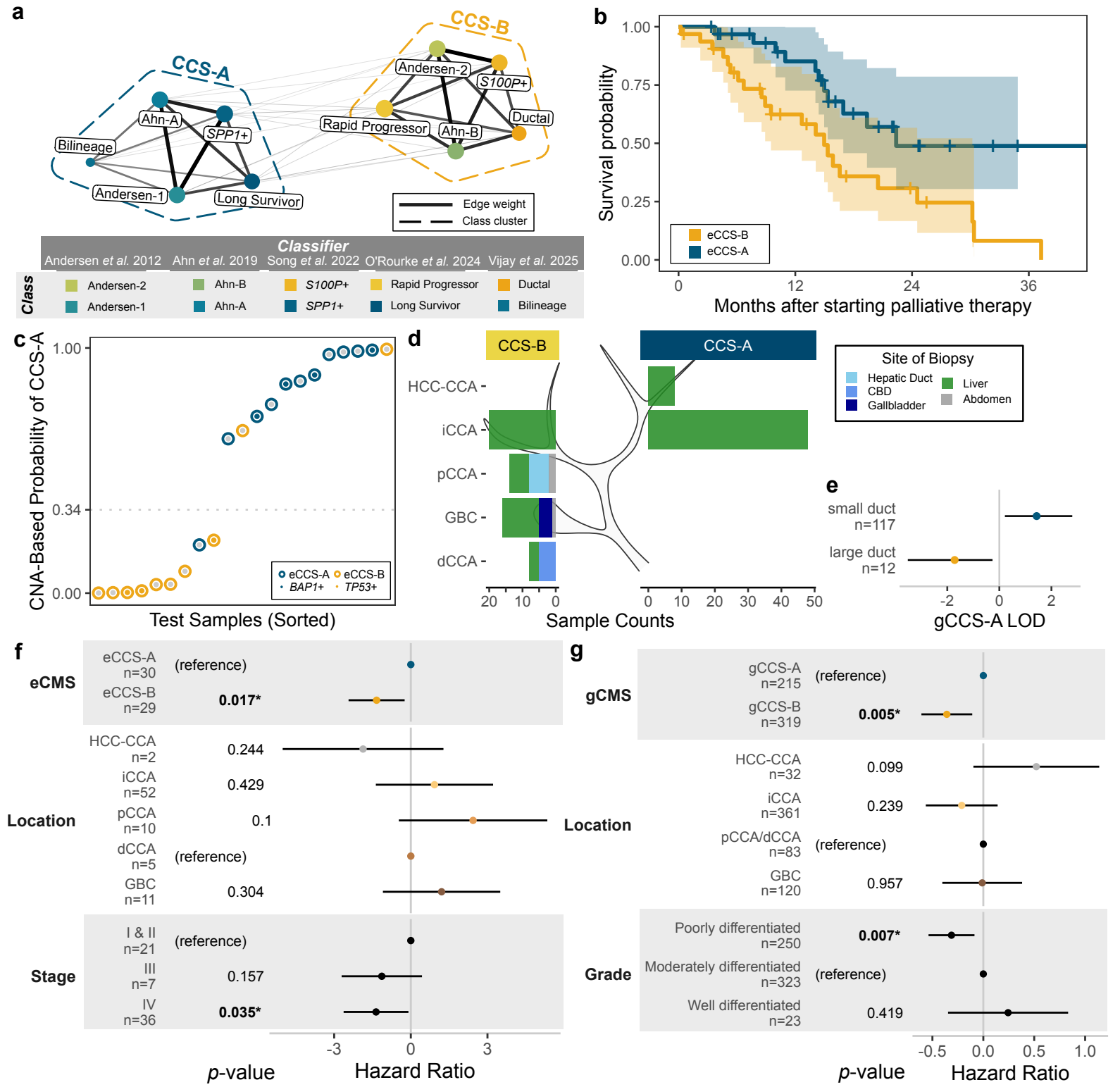
a
CCS-A
CCS-B
Ahn-A
SPP1+
Bilineage
Andersen-1
Long Survivor
Andersen-2
S100P+
Rapid Progressor
Ahn-B
Ductal
Edge weight
Class cluster
Classifier
Andersen et al. 2012
Ahn et al. 2019
Song et al. 2022
O'Rourke et al. 2024
Vijay et al. 2025
Class
Andersen-2
Andersen-1
Ahn-B
Ahn-A
S100P+
SPP1+
Rapid Progressor
Long Survivor
Ductal
Bilineage
b
Survival probability
1.00
0.75
0.50
0.25
0.00
eCCS-B
eCCS-A
0
12
24
36
Months after starting palliative therapy
c
CNA-Based Probability of CCS-A
1.00
0.34
0.00
eCCS-A
eCCS-B
BAP1+
TP53+
Test Samples (Sorted)
d
CCS-B
CCS-A
HCC-CCA
iCCA
pCCA
GBC
dCCA
20
10
0
0
10
20
30
40
50
Sample Counts
Site of Biopsy
Hepatic Duct
CBD
Gallbladder
Liver
Abdomen
e
small duct
n=117
large duct
n=12
-2
0
2
gCCS-A LOD
f
eCMS
eCCS-A
n=30
(reference)
eCCS-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*
-3
0
3
p-value
Hazard Ratio
g
gCMS
gCCS-A
n=215
(reference)
gCCS-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
-0.5
0.0
0.5
1.0
p-value
Hazard Ratio

Figure 2

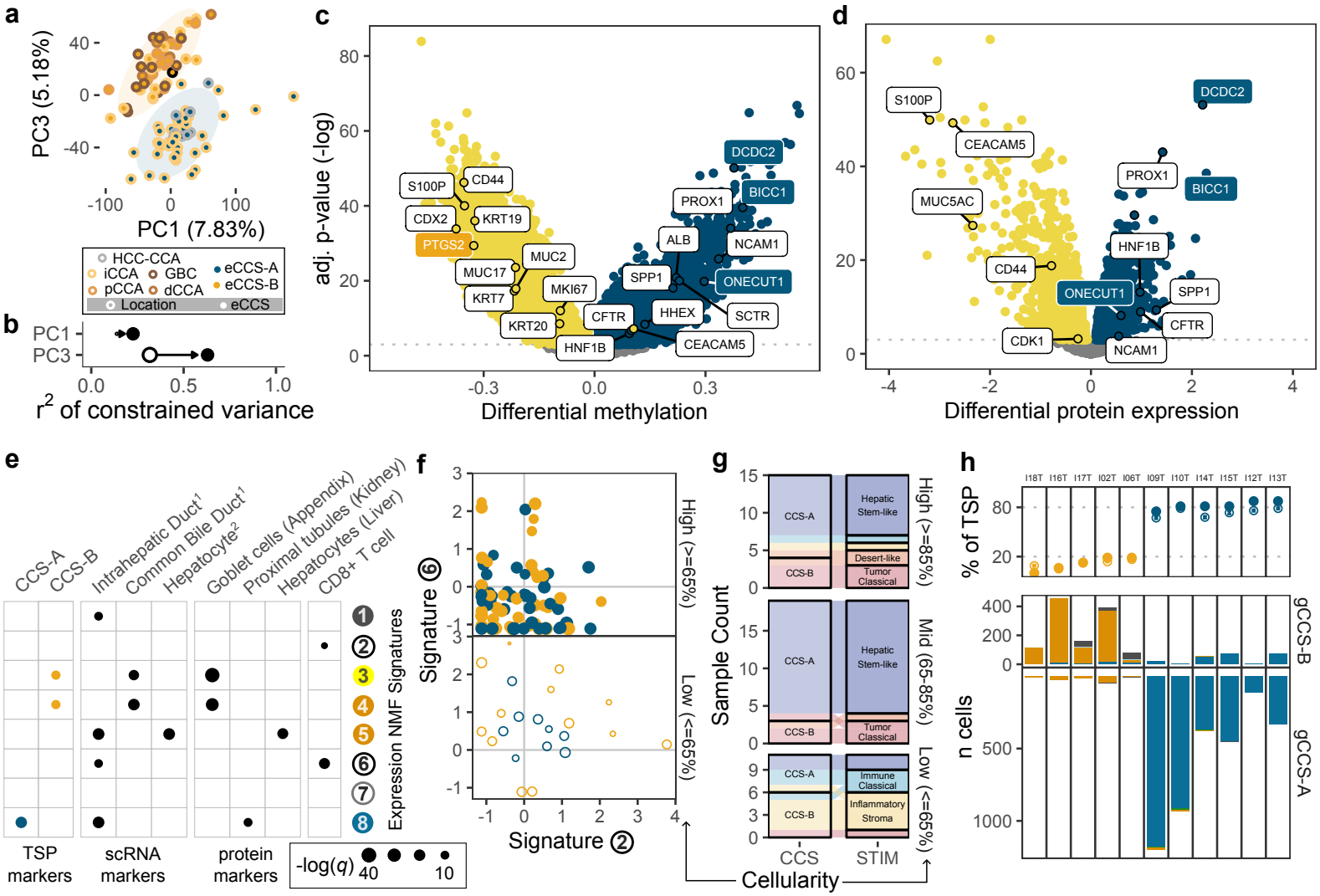
a
PC3 (5.18%)
PC1 (7.83%)
HCC-CCA
iCCA
GBC
pCCA
dCCA
eCCS-A
eCCS-B
Location
eCCS
b
PC1
PC3
r2 of constrained variance
c
adj. p-value (-log)
Differential methylation
S100P
CD44
CDX2
KRT19
PTGS2
MUC2
MUC17
KRT7
MKI67
KRT20
HNF1B
CFTR
HHEX
CEACAM5
SCTR
SPP1
ALB
NCAM1
ONECUT1
PROX1
BICC1
DCDC2
d
Differential protein expression
S100P
CEACAM5
MUC5AC
CD44
CDK1
ONECUT1
HNF1B
PROX1
BICC1
DCDC2
SPP1
CFTR
NCAM1
e
CCS-A
CCS-B
Intrahepatic Duct1
Common Bile Duct1
Hepatocyte2
Goblet cells (Appendix)
Proximal tubules (Kidney)
Hepatocytes (Liver)
CD8+ T cell
Expression NMF Signatures
TSP markers
scRNA markers
protein markers
-log(q)
40
10
f
Signature ⑥
Signature ②
High (>=65%)
Low (<=65%)
g
Sample Count
CCS-A
CCS-B
Hepatic Stem-like
Tumor Classical
Desert-like
Mid (65-85%)
High (>=85%)
Low (<=65%)
Immune Classical
Inflammatory Stroma
CCS
STIM
Cellularity
h
% of TSP
n cells
I18T
I16T
I17T
I02T
I06T
I09T
I10T
I14T
I15T
I12T
I13T
gCCS-B
gCCS-A

Figure 3

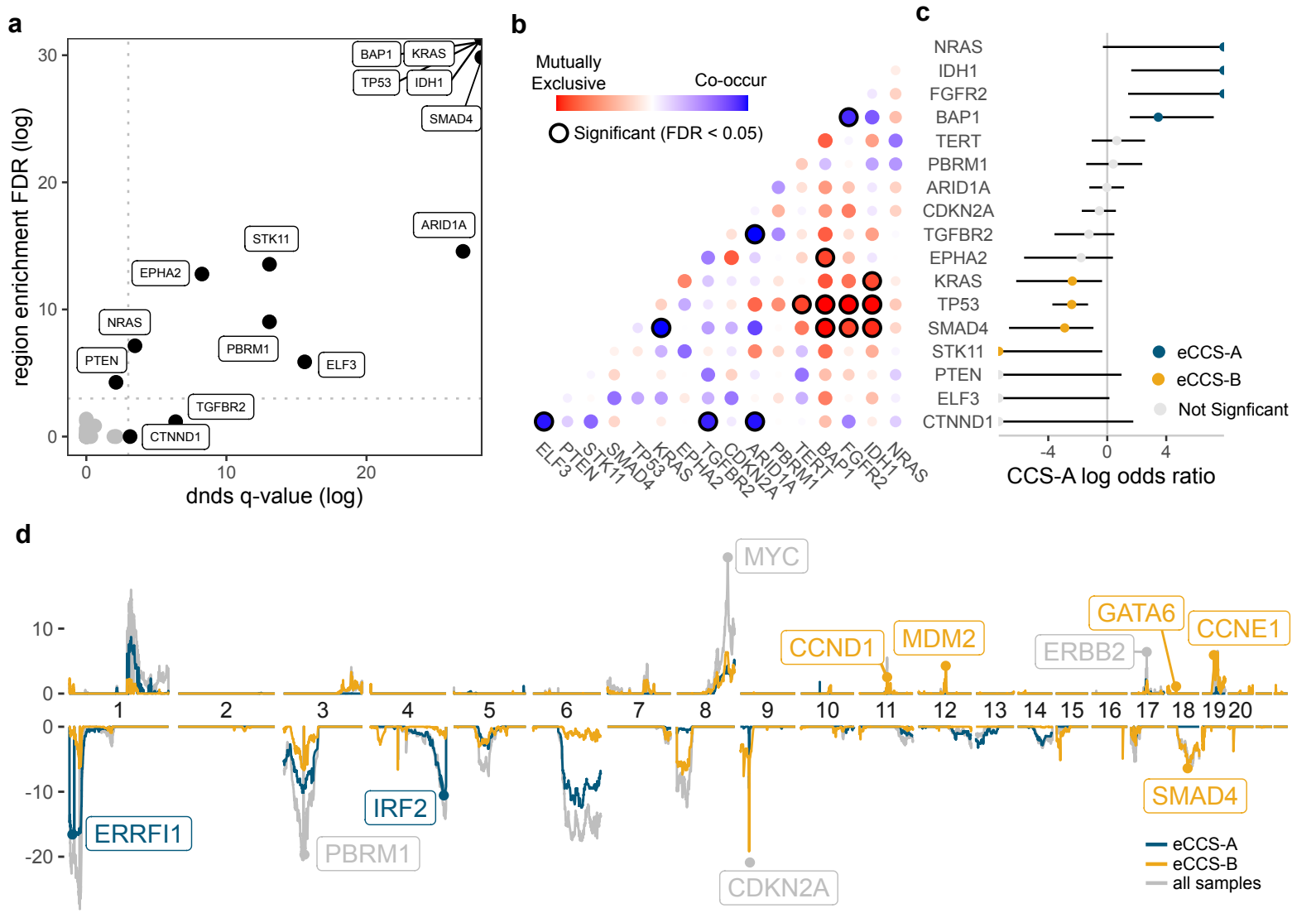
a
region enrichment FDR (log)
dnds q-value (log)
BAP1
KRAS
TP53
IDH1
SMAD4
ARID1A
STK11
EPHA2
NRAS
PBRM1
ELF3
PTEN
TGFBR2
CTNND1
b
Mutually Exclusive
Co-occur
Significant (FDR < 0.05)
ELF3
PTEN
STK11
SMAD4
TP53
KRAS
EPHA2
TGFBR2
CDKN2A
ARID1A
PBRM1
TERT
BAP1
FGFR2
IDH1
NRAS
c
NRAS
IDH1
FGFR2
BAP1
TERT
PBRM1
ARID1A
CDKN2A
TGFBR2
EPHA2
KRAS
TP53
SMAD4
STK11
PTEN
ELF3
CTNND1
eCCS-A
eCCS-B
Not Signficant
CCS-A log odds ratio
d
MYC
CCND1
MDM2
ERBB2
GATA6
CCNE1
ERRFI1
IRF2
PBRM1
CDKN2A
SMAD4
eCCS-A
eCCS-B
all samples

Figure 4

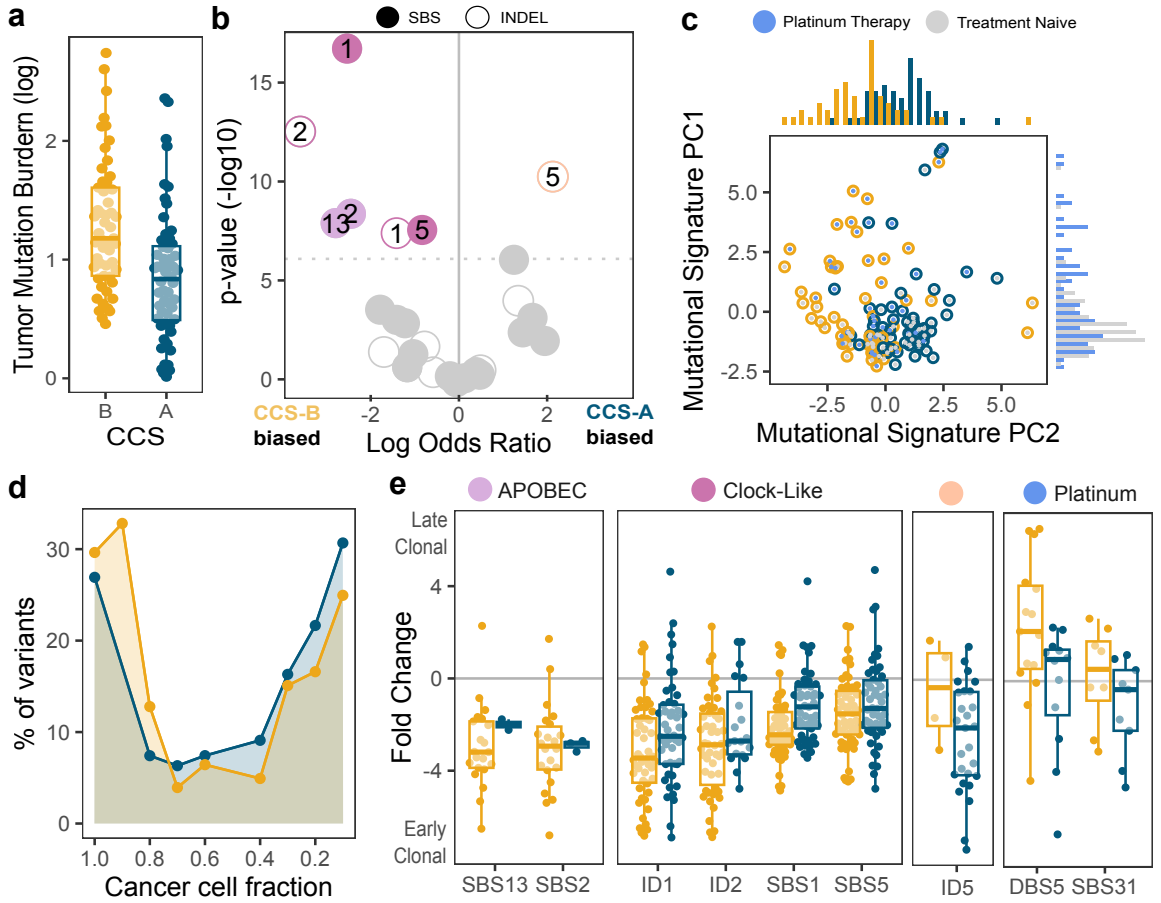
a
Tumor Mutation Burdern (log)
CCS
B
A
b
SBS
INDEL
p-value (-log10)
Log Odds Ratio
CCS-B biased
CCS-A biased
c
Platinum Therapy
Treatment Naive
Mutational Signature PC1
Mutational Signature PC2
d
% of variants
Cancer cell fraction
e
APOBEC
Clock-Like
Platinum
Fold Change
Late Clonal
Early Clonal
SBS13
SBS2
ID1
ID2
SBS1
SBS5
ID5
DBS5
SBS31

Figure 5