Genomic characterization of co-existing neoplasia and carcinoma lesions reveals distinct evolutionary paths of gallbladder cancer

Jianzhen Lin1,2,7, Xinxin Peng3,7, Kun Dong4,7, Junyu Long1,7, Xuejiao Guo3, Hongyue Li3, Yi Bai1, Xu Yang1, Dongxu Wang1, Xin Lu1, Yilei Mao1, Xinting Sang1, Xuwo Ji3, Haitao Zhao1✉ & Han Liang5,6✉

Gallbladder carcinoma is the most common cancer of the biliary tract with dismal survival largely due to delayed diagnosis. Biliary tract intraepithelial neoplasia (BilIN) is the common benign tumor that is suspected to be precancerous lesions. However, the genetic and evolutionary relationships between BilIN and carcinoma remain unclear. Here we perform whole-exome sequencing of coexisting low-grade BilIN (adenoma), high-grade BilIN, and carcinoma lesions, and normal tissues from the same patients. We identify aging as a major factor contributing to accumulated mutations and a critical role of \textit{CTNNB1} mutations in these tumors. We reveal two distinct carcinoma evolutionary paths: carcinoma can either diverge earlier and evolve more independently or form through the classic adenoma/dysplasia-carcinoma sequence model. Our analysis suggests that extensive loss-of-heterozygosity and mutation events in the initial stage tend to result in a cancerous niche, leading to the subsequent BilIN-independent path. These results reframes our understanding of tumor transformation and the evolutionary trajectory of carcinogenesis in the gallbladder, laying a foundation for the early diagnosis and effective treatment of gallbladder cancer.
Gallbladder carcinoma (GBC) is the most common cancer of the biliary tract. Its incidence rate shows a wide geographic variance, and in some countries such as Chile and India, the incidence rate is nearly 30/100000 in women. Unlike other cancer types, only 30% of GBC patients are diagnosed or suspected preoperatively, whereas other cases are diagnosed based on postoperative incidental findings. In most patients, the tumor is recognized either at the time of surgery or by the pathologist examining the surgical specimen. Most symptomatic GBC patients have an incurable tumor, and the risk of GBC recurrence is 64% at 5 years, even after R0 resection. The clinical outcome of GBC is very poor: although a 5-year survival rate of 75% can be achieved in early-stage (T1) disease, the overall 5-year survival rate is <5%. Therefore, there is an urgent clinical need to identify high-risk patients and have their gallbladders removed before the development of GBC.

When considering the potential risk for gallbladder cancer, it is important to understand the relationship between gallbladder polypoid lesions and GBC, thereby managing them wisely and timely. The proportion prevalence of gallbladder polyps is ~5%, accounting for 2–12% of the cholecystectomy specimens. Low-grade biliary tract intraepithelial neoplasia (LG-BilIN) and high-grade biliary tract intraepithelial neoplasia (HG-BilIN), two common benign neoplastic polyps, have attracted much attention. Additionally, we separated GBC, LG-BilIN, HG-BilIN, and normal gallbladder tissues and performed whole-exome sequencing on 44 samples in total (mean depth: normal 107×; GBC 171×; LG-BilIN 188×; HG-BilIN 206×) (Fig. 1a and Supplementary Data 1). We employed a well-established, multi-caller-based MC3 approach to call single nucleotide variant (SNV) mutations and small indels, and this approach has been used to generate the high-quality mutation data of The Cancer Genome Atlas (TCGA). Importantly, we performed a rescue step to call mutations across different samples from the same patient to reduce false negatives. Across the 33 tumor samples from the 11 patients, we identified an average of 214 SNVs (range: 69–725) and 9 indels (range: 3–48) (Supplementary Fig. 1a and Supplementary Data 2). To assess the accuracy of our mutation calls, we performed independent Sanger sequencing and reached a validation rate of 95% (42 out of 44 mutations) (Supplementary Fig. 2 and Supplementary Data 3). To assess the relative mutation burden of GBC, we compared our GBC samples with stage-I samples of 16 other cancer types from TCGA and found that GBC exhibited a moderate level of tumor mutation burden (TMB), similar to that of liver hepatocellular carcinoma (LIHC) (t-test, p = 0.79) but significantly higher than that of cholangiocarcinoma (CHOL) (t-test, p = 7.3 × 10^{-4}, Fig. 1b).

In order to identify the mutagenesis mechanisms underlying the mutations and evolution of these GBC samples and elucidate the common/specific contributing factors, we decomposed the mutational spectrum for the 16 TCGA cancer types (stage-I only) as well as our GBC samples, based on 30 established mutational signatures (COSMIC mutational signature v2)\(^3\). This analysis confirmed some known signature patterns such as smoking (Signature 4) for lung cancers (Fig. 1c). We found that in most cancer types, the mutations in stage-I tumors were mainly caused by age, especially for colorectal cancer (COAD/READ), pancreatic adenocarcinoma (PAAD), and stomach cancer (STAD). In the GBC samples, age-related Signature 1 contributed to >75% of the mutations and APOBEC-related Signature 2 contributed to ~10% (Fig. 1c). These results indicate the importance of mutations accumulated by the aging process for GBC development and provide the first view of mutational characteristics of GBC coexisting with LG/HG-BilIN.

Somatic alteration feature comparison of gallbladder tumors. In addition to the above-mentioned comparison among malignant tumor types, we compared the mutational features among the three gallbladder tumor types. We found that GBC, LG-BilIN, and HG-BilIN showed similar mutation rates (paired t-test, p > 0.05, Fig. 2a), but the tumor samples from the two young patients (P03 and P11) showed much fewer mutations than those from older patients, regardless of tumor types (mean number, 101 vs. 273, t-test, p < 7 × 10^{-3}, Fig. 2b). We further quantified the microsatellite instability (MSI) status of these tumor samples based on a large set...
Fig. 1 Study overview and mutation characteristics of BilIN-related GBC. a Overall study design. b The TMB distribution of BilIN-related gallbladder carcinoma (GBC) in the context of stage-I samples from 16 TCGA cancer types. c Relative contributions of different mutational signatures (annotated by etiology factors) in GBC and 16 TCGA cancer types. b, c GBC is highlighted in the red box. READ: rectum adenocarcinoma; THCA: thyroid carcinoma; TGCT: testicular germ cell tumors; KICH: kidney chromophobe; BRCA: breast invasive carcinoma; CHOL: cholangiocarcinoma; PAAD: pancreatic adenocarcinoma; KIRC: kidney renal cell carcinoma; KIRP: kidney renal papillary cell carcinoma; LIHC: liver hepatocellular carcinoma; HNSC: head and neck squamous cell carcinoma; ESCA: esophageal carcinoma; COAD: colon adenocarcinoma; STAD: stomach adenocarcinoma; LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma.
Fig. 2 Mutational feature comparison among three types of gallbladder tumors. **a** TMB distribution per tumor type (n = 11 samples). **b** The number of mutations in tumor samples from young (n = 6 samples) and old patients (n = 27 samples). **c** MSI score distribution per tumor type (n = 10 samples). **d** Relative contributions of different mutational signatures (annotated by etiology factors) in tumor samples based on COSMIC mutational signature v2. **e** An oncoplot showing potential driver mutations detected in at least two patients. Mutations are colored according to different substitution types. **a-c** The middle line in the box is the median, the bottom, and top of the box are the first and third quartiles, and the whiskers extend to 1.5× interquartile range of the lower and the upper quartiles, respectively. Raw p-values based on two-tailed Student’s t-test are shown.
of selective microsatellite loci24 but did not find a significant difference among tumor types (paired t-test, p > 0.05, Fig. 2c). The three tumor types also showed similar compositions in terms of six possible substitution classes (Supplementary Fig. 1b). Based on COSMIC mutational signature v2, different tumors from the same patient generally showed similar mutation signature patterns, but the signature composition varied considerably from patient to patient (Fig. 2d), suggesting that the underlying mutagenesis mechanisms are more patient-specific than tumor type-specific. Across the 33 tumor samples, there was a consistent component of age-related Signature 1. In addition, tumor samples from patient P10 showed APOBEC-related Signature 2 and ultraviolet exposure related Signature 7; tumors from patient P05 exhibited APOBEC-related signatures (Signature 2 and Signature 13) and Signature 7 (although Signature 7 is found predominantly in skin, head and neck or oral squamous cancers, it has been identified in other cancer types25). Tumors from P02 showed a sizable contribution from POLE-related Signature 10, which was further confirmed by the mutational analysis based on the COMSIC mutation signature v3 (Supplementary Fig. 3). The GBC samples from P03, P06, and P08 showed some signature divergences from their corresponding benign tumors, LG-BilIN, and HG-BilIN.

To understand key somatic mutations during tumor development, we examined the potential somatic mutation drivers. To overcome the relatively small sample size of our cohort, we focused on known drivers identified in a recent TCGA pan-cancer analysis26. In total, there were 16 potential mutation drivers observed in at least two patients. Among them, mutations in CTNNB1 and ARID2 were detected in five out of the 11 patients, and TP53 and ERBB3 mutations were identified in three patients (Fig. 2e). GBC samples did not contain more driver mutations than LG- or HG-BilIN. As for specific mutations, five patients harbored four different CTNNB1 mutations, including p.T411I shared by all three tumor samples from P03, p.S45F shared in P07 and P10, p.S33C in P09, and p.S45F and p.K335T in LG-BilIN and GBC from P01, respectively. All these CTNNB1 mutations were probably damaging, as suggested by a PolyPhen2 score of >0.9. CTNNB1 mutations have been reported to cause modified β-catenin activity associated with liver tumor progression28. This observation highlights the key role of β-catenin activity driven by CTNNB1 mutations for cellular transformation in GBC tumorigenesis. For mutations in ARID2, a subunit of the PBAF chromatin-remodeling complex, five patients harbored six mutations: only p.Q916* was shared by the three tumor samples from P10, suggesting loss-of-function ARID2 as a driver in this patient; in the other four patients, ARID2 mutations were only detected in GBC or BilIN samples. Among the three patients harboring ERBB3 mutations, one mutation was shared in all three samples of P10, and the other two patients (P04 and P06) only contained a mutation in BilIN samples. We observed SMARCA4 mutations in the GBC samples but not in BilIN samples from P04 and P11, suggesting that those mutations were acquired in a late stage in GBC development. We also investigated the somatic copy-number alterations (SCNAS) inferred from WES data across the 33 tumor samples. Among the 11 GBC samples, we noted significant copy-number gain of ERBB2 (four samples) and significant copy-number loss of CDKN2A (four samples). These SCNAS may have played a role in the GBC tumorigenesis (Supplementary Fig. 4). Altogether, these results provide a shortlist of driver somatic alteration candidates in the development of adenoma/dysplasia-related GBC.

**Two GBC evolutionary paths driven by early LOH and mutation loads.** Somatic mutations provide a molecular footprint for cancer evolution. Considering the high reliability for SNV somatic mutations, we inferred the evolutionary relationships among GBC, LG-, and HG-BilIN based on their SNV mutations primarily. To ensure the accuracy of the evolutionary trees, we employed two independent methods, Treeomics29 and MEGAX30, and observed very consistent tumor sample phylogenetic topology for all 11 patients (Methods). The trees could be classified into two groups: (i) the BilIN-independent group (6 patients) in which GBC split before the common ancestor of LG-BilIN and HG-BilIN, and GBC evolved more independently from the two BilIN tumors (Fig. 3a); and (ii) the BilIN-dependent group (5 patients) in which GBC split after the common ancestor of LG-BilIN and HG-BilIN, and GBC was clustered more closely with HG-BilIN in two patients and with LG-BilIN in three patients (Fig. 3b). To assess the classification robustness, we performed the same analysis on the phylogenetic trees inferred from gene-level SCNA values or large-scale loss-of-heterozygosity (LOH) status, respectively, and found that the mutation-tree-group classification indeed represent the consensus assignment in all the cases (Supplementary Data 4). Interestingly, the two young patients (P03 and P11, diagnosed at 38 years) were classified into separate groups (P03: BilIN-dependent; P11: BilIN-independent). Specifically, GBC, LG-BilIN, and HG-BilIN from P03 only shared one potential driver mutation (CTNNB1.PT411), while GBC in P11 had one SMARCA4 mutation, and HG-BilIN in P11 harbored one ARID2 mutation. These results suggest two coexisting evolutionary paths for GBC development: the BilIN-dependent model is more aligned with the traditional adenoma/dysplasia-carcinoma sequence model, wherein benign tumors, LG-/HG-BilIN, serve as the precursors of GBC, and the BilIN-independent model that suggests an earlier diverged origin of GBC. Interestingly, although both the number of mutations and the age at diagnosis were very similar in the two groups (Fig. 3c, d), the number of mutations acquired in the common ancestor of LG-BilIN/HG-BilIN/GBC was significantly higher in the BilIN-independent group than in the BilIN-dependent group (mean, 150 vs. 36, Wilcoxon rank-sum test, p <0.05). These results suggest that the mutation load accumulated in the early stage of tumor development, rather than the total mutation load in GBC, affects the evolutionary path of GBC.

To identify other driving factors affecting the evolutionary paths of GBC, we compared genomic features of GBC samples between the two groups but did not detect any significant differences in specific mutated drivers and MSL. LOH is a common occurrence in cancer, and the accumulation of LOH can be associated with carcinogenesis of the gallbladder31,32. Interestingly, we observed that GBC samples in the BilIN-independent group showed a higher proportion of LOH than those in the BilIN-dependent group (mean, 20% vs. 5%, t-test, p = 0.021, Fig. 4a). This observation raised two possible evolutionary scenarios for the BilIN-independent group: (i) large-scale LOH events occurred in the common ancestor of LG-BilIN/HG-BilIN/GBC, which then led to an early split of GBC; or (ii) after the divergence from LG-BilIN/HG-BilIN, LOH events independently occurred in the GBC lineage. To identify the more likely scenario, we compared the LOH status of LG-BilIN/HG-BilIN and found that they showed the same pattern as GBC samples (mean, 20% vs. 3%, t-test, p = 1.8 × 10^-4, Fig. 4b). We further compared the LOH status of the common ancestor of LG-BilIN/HG-BilIN/GBC (by considering LOH regions shared among the three tumors) and observed that LOH events were significantly more extensive in the BilIN-independent group than in the BilIN-dependent group (mean, 13% vs. 2%, t-test, p = 0.028, Fig. 4c). Finally, to rule out the possibility that LOH regions may affect the mutation-based tree construction, we rebuilt the phylogenetic trees based only on the mutations in non-LOH regions and observed the same patterns (Supplementary Fig. 5). When examining the LOH status of the common ancestor of LG-BilIN/
**Fig. 3 Phylogenetic trees of gallbladder tumors.**

**a** BilIN-independent group (n = 6), and **b** BilIN-dependent group (n = 5). The branch lengths are proportional to the number of supporting mutations, which were labeled. Potential driver mutations are annotated along the branches. Bootstrap values are shown in orange.

**c** The number of mutations in GBC samples (BilIN-independent: n = 6; BilIN-dependent: n = 5), **d** age at diagnosis (BilIN-independent: n = 6; BilIN-dependent: n = 5), and **e** the mutations accumulated in the common ancestor state of LG-BilIN/HG-BilIN/GBC, in the BilIN-independent and BilIN-dependent groups (BilIN-independent: n = 6; BilIN-dependent: n = 5). The middle line in the box is the median, the bottom, and top of the box are the first and third quartiles, and the whiskers extend to 1.5× interquartile range of the lower and the upper quartiles, respectively. Raw p-values based on the two-tailed Wilcoxon rank-sum test are shown.
HG-BilIN/GBC samples at the chromosome level, we found that chr5, chr17, chr18, and chr20 showed the most striking differences (t-test, \( p < 0.05 \), Fig. 4d). Indeed, the clustering pattern of gallbladder tumors based on their chromosome-level LOH status largely recaptured the tumor group classification based on mutation-based tree topologies (Fig. 4e). To gain more insights into the clonal evolution, we estimated the purity and ploidy of GBC/LG-BilIN/HG-BilIN samples using FACETS\(^\text{33}\) and then inferred the number of clones for tumor evolution in each patient using PyClone\(^\text{34}\). We found that tumors in the BilIN-independent group contained significantly more clones than those in the BilIN-dependent group (t-test, \( p = 0.049 \), Fig. 4f), although they did not
contain more somatic mutations (t-test, p = 0.76, Fig. 4g). We also investigated the possible seeding patterns for GBC, based on inferred cancer cell fractions (CCF) of mutations and found that all 11 GBC cases were polyclonal seeding (Supplementary Data 5).

Based on these results, we proposed an evolutionary model of GBC development in which the extent of LOH events and somatic mutations accumulated at the initial stage of gallbladder tumor development is the key to determine the subsequent path of carcinoma evolution (Fig. 5). In the BilIN-independent group, extensive LOH and mutation events occur first, leading to a “cancerous” niche in the common ancestor of LG-BilIN/HG-BilIN/GBC. Given the acquired cancer potentiality of the cancerous niche, GBC then tends to split earlier and evolve more independently from LG-BilIN and HG-BilIN, and there are more clones in the microevolutionary process. In the BilIN-dependent group, fewer LOH events and somatic mutations occur initially, resulting in a “neoplastic” niche in the common ancestor. The neoplastic niche needs a step-wise progression to achieve the malignant state. Rare. Like most epithelial cancer types, GBC is preceded by a series of precancerous histological and molecular changes that evolve possibly over a period of several decades. Adenoma and hyperplastic changes progressing to dysplasia are often assumed to be the usual form of GBC, offering the opportunity to elucidate cancer pathogenesis and natural history in this disease. Through a genome-wide characterization of somatic alterations in a series of coexisting LG-BilIN, HG-BilIN, and GBC, we have deciphered a detailed molecular portrait of how GBC evolves in the context of gallbladder polyps. The simultaneous analysis of different gallbladder tumor types (both benign and malignant) from the same patients allowed us to reconstruct the evolutionary history with high confidence, distinguishing our study from previous investigations on GBC that were based on comparisons between precancerous and carcinomatous lesions from separate patient populations.

Our mutational analysis highlights that all LG-BilIN, HG-BilIN, and GBC are an inevitable consequence of aging and mutation accumulation, characterized by an age-related mutational signature (S1) as a major component observed in all the tumors. Putting this finding in the context of the median age of onset of these pathomorphological changes is informative. Gallbladder polyps with symptomatic cholecytitis seldom appear before the age of 40, the median age for detection of adenoma (LG-BilIN) is 45 years, and that for HG-BilIN is 50 years, while the median age of GBC diagnosis is 71 years. It can thus be
speculated that the cancerous lesions start as small clusters of neoplastic epithelial cells at an early stage that then increases in size and number over the years. At the time of cancerous transition, mutation-driven uncontrollable proliferation usually selects the clones with oncogenic potential. We observed several recurrent driver mutations in GBC, including TP53, KRAS, ARID2, SMAD4, PI3KCA, and ERBB3, similar to a previous study38. Notably, our results suggest a critical role of CTNNBI mutations in the cellular transformation from normal cell growth into a state of uncontrollable proliferation. Previous studies reported that adenoma or LG-BilIN expresses significantly higher β-catenin than GBC, and 62.5% of LG-BilIN samples harbored mutations in CTNNBI exon 3, but this rate was only 4.8% in the GBC39,40. In our cohort, 5 of the 11 patients harbored CTNNBI mutations; in 4 of these 5 cases, GBC and its adjacent LG-BilIN shared the same mutations in CTNNBI, suggesting that CTNNBI mutations play a more active role in the formation of carcinoma in the context of adenoma/dysplasia than GBC alone.

Two major developmental paths have been suggested for GBC: (i) de novo development when only GBC is present; and (ii) the adenoma/dysplasia–carcinoma sequence model when GBC coexisted with LG-BilIN or HG-BilIN2. However, it remains unclear how the GBC developmental paths are determined and whether LG-BilIN or HG-BilIN represents a major transition state in the latter model. Through a comprehensive, multilayer analysis of the cancer genome, our study suggests that the relationship between GBC and adjacent LG-BilIN/HG-BilIN can be divided into two evolutionary paths, the BilIN-independent path, and the BilIN-dependent path. The BilIN-independent model is more aligned with the de novo development path wherein the carcinoma initiates before the divergence of LG-BilIN and HG-BilIN, and evolves more independently. In contrast, the BilIN-dependent divergence model is similar to the adenoma/dysplasia–carcinoma sequence where carcinoma evolves from LG-BilIN or HG-BilIN through a step-wise progressive manner. The key determinant is the accumulation of LOH and somatic mutation events in the initial stage that is sufficient to form a cancerous niche for subsequent GBC development. Interestingly, besides somatic mutations, we uncovered contrasting patterns of LOH events between the two evolutionary paths, which highlights the importance of losing functional tumor suppressor genes in the niche-defining process. Among individual chromosomes enriched for LOH events that presumably promote the cancerous niche, LOH on chr.5q and chr.7p have been associated with early carcinogenic changes of the gallbladder31, and importantly, LOH events are rare in the normal gallbladder22. Clinically, our results suggest that proactive intervention needs to focus on identifying patients with LOH-rich gallbladder tissue through noninvasive detection methods, so that earlier or more aggressive cholecystectomy may be adopted as a primary prevention approach for GBC. Furthermore, our results confirm that both dysplasia–carcinoma sequence (HG-BilIN-GBC) and adenoma–carcinoma sequence (LG-BilIN-GBC) exist in the GBC development. We did not observe significant bias towards LG-BilIN or HG-BilIN, despite the small sample size of our BilIN-dependent group (HG-BilIN-GBC, n = 2; LG-BilIN-GBC, n = 3). It remains unclear whether residual dysplasia at the cystic duct margin after cholecystectomy is associated with an increased incidence of recurrent carcinoma31–33. Thus, timely detection of GBC with neoplastic niches is critical in deciding the need for a BilIN-free surgical margin.

One limitation of this study is the representativeness of our cases. Since the cases (carcinomas coexisting with both adenoma and BilIN) surveyed are rare among gallbladder cancer with a prevalence from 1.8% to 6.4%36,44,45, further efforts are required to examine the evolutionary paths of more common gallbladder cancer. Other limitations may be the relatively small cohort size and the selection of manual sampling locations. To mitigate these, we have made extensive efforts in searching for such cases from a large tumor tissue bank, and have performed a careful pathological examination of all the sections of the resected gallbladder. We also ensured that every normal control sample was tumor-, adenoma-, and dysplasia-free mucosa, thereby removing the potential bias in phylogenetic tree inference. The mutations in cancer driver genes identified in our cohort are similar to previous findings, supporting their representativeness39,48. Further efforts are needed to confirm our evolutionary model using independent cohorts, especially prospective studies with sequential samplings from the normal gallbladder, LG-BilIN, emerging HG-BilIN, and GBC formation, over the disease course. The mechanisms through which different LOH events and mutation loads (high vs. low) could drive normal gallbladder mucosa into distinct states of gallbladder cancerous/neoplastic “niche” remain unclear. Further multi-omics characterization of epigenetic alterations, aneuploidy, or transcriptomes will provide more insights into these oncogenic transformations.

Methods

Patient recruitment and sample cohort: All samples were collected with the approval of the Institutional Review Board (IRB) from Peking Union Medical College Hospital (PUMCH), Beijing, China. Informed and written consent was obtained from each patient. We enrolled 11 patients who underwent surgical resection of gallbladder polyps at PUMCH from April 2008 to December 2016 and had pathologically confirmed geographically coexisting GBC, LG-BilIN, and HG-BilIN lesions, according to the 2019 WHO classification of tumors of the digestive system36. For LG-BilIN, we only included gallbladder adenoma with low-grade dysplasia, referring to the International Classification of Diseases for Oncology (ICD-O) code as 8303/0. The HG-BilIN indicated the regions with high-grade dysplasia or carcinoma in situ (CIS), referring to ICD-O code as 8530/3. All GBC lesions were T1 stage tumors, including tumors invaded lamina propria and tumors infiltrating muscle layer. The inclusion of samples required GBC, LG-BilIN, HG-BilIN, and normal gallbladder tissues for each patient. Independent pathologists evaluated the pathology of all the samples in this study and obtained consistent results. To minimize the cancerization effect on the normal samples, we only included normal gallbladder tissues in a paraffin block that contained no cancer. All clinical data were obtained from the hospital records.

Whole-exome sequencing: Serial consecutive 5-μm-thick sections were cut from representative deparaffinized blocks and stained with hematoxylin. Multiple pathologists independently confirmed the diagnosis, different histological features, and adequate cellularity. Four components, GBC, LG-BilIN, HG-BilIN, and normal mucosa, were separately laser-microdissected using a Leica CTR 6000 Microsystem (Wetzlar, Germany) from the slides of each patient. Genomic DNA (gDNA) was extracted using the QIAamp DNA FFPE Tissue Kit (Qiagen, Germany) according to the manufacturer’s protocol. The detailed DNA fragments were repaired and 3′-tailed. Adaptors were ligated to both ends of the fragments, followed by size selection. Size-selected fragments were amplified via polymerase chain reaction (PCR). Exome capture was performed using SureSelect Human All Exon V6 (Agilent) according to the manufacturer’s protocol, followed by PCR amplification. Libraries were sequenced on Illumina NovaSeq.

Analysis of somatic mutations: Whole-exome sequencing read pairs were trimmed, and only read pairs with <3% N bases and >50% high-quality bases were kept for subsequent analyses. The resulting high-quality reads were aligned to the human reference genome (Homome_sapiens_assembly19) using Burrows-Wheeler Aligner (0.7.17)47. To improve the alignment accuracy, we used Genome Analysis Toolkit (GATK, version 3.8.1)48 to process BAM files, including marking duplicates, base quality recalibration, and local realignment around high-confidence insertions and deletions. Based on ~7000 high-frequency SNP sites, the identical marking dupli-

NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-021-25012-9 | www.nature.com/naturecommunications

NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-021-25012-9 | www.nature.com/naturecommunications
nonsense mutations or ORF-shift mutations in the overlapped targeted regions in this study with those defined in the TCGA MC3 project. For the GBC/LG-BILIN/ HG-BILIN analysis, we further reconciled their mutation calls from the three samples of the same patient to increase the sensitivity of mutation detection. Specifically, for a specific mutation identified in only one or two samples in a patient, we rescued it in the remaining sample(s) of the same patient even if only one caller supported the mutation. We used MANTIS24 to call the MSI patient, we rescued it in the remaining sample(s) of the same patient even if the HG-BilIN analysis, we further reconciled their mutation calls from the three patients. To infer the tumor seeding pattern for GBC patients, we compared the genomic regions. We also performed the phylogenetic tree reconstruction based on gene-phylogenetic tree when including mutations detected outside the exome-targeted regions. We also performed the phylogenetic tree reconstruction based on gene-level SNVs and large-scale LOH status as described above, respectively (Supplementary Data 4). According to the phylogenetic trees, we classified the 11 patients into two groups based on whether GBC split before or after the common ancestor of LG-BILIN and HG-BILIN: one group was labeled as BILIN-independent divergence (six patients), and the other was labeled as BILIN-dependent divergence (five patients). To remove the potential bias of LOH regions in mutation calling, we repeated the tree constructions using only the mutations in the non-LOH regions.

Analysis of clonal evolution. We employed PyClone34, a statistical model based on a Bayesian clustering method, to infer the cancer cell fractions (CCF) for each SNV mutation with a Beta Binomial emission. The purity and ploidy for most samples were successfully estimated by FACETS based on segmented genome35 except for those from P03 and P07, whose genome showed few alterations. The samples from these two patients were further estimated by Control-FREEC36 with a ploidy set to 2. For each SNV mutation, we extracted its major and minor copy numbers generated by FACETS. Based on PyClone estimation, we only counted those clones (mutation clusters) with ≥10 mutations. To infer the tumor seeding pattern for GBC patients, we compared the CCFs between sample pairs and calculated the Jaccard similarity index (JSI) for each pair as previously described35 and the JSI cutoff of 0.3 to distinguish monoclonal versus polyclonal seeding.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The raw WES data used in this study are available in the Genome Sequence Archive (GSA) database under accession code HRA000029 and the European Genome-Phenome Archive (EGA) database under the accession code EGAS0000105402. The person who wants to access the data can contact The Data Access Committee (DAC) listed at the websites. The access applications should be sent by email with a detailed introduction on the research intended to carry out based on the data. Somatic mutation data are also available in the European Variation Archive (EVA) database under the accession code PRED84269. Source data are available as a Source Data file. The remaining data are available within the Article and Supplementary Information. Source data are provided with this paper.

Received: 31 July 2020; Accepted: 16 July 2021; Published online: 06 August 2021

References

1. Varshney, S., Butturini, G. & Gupta, R. Incidental carcinoma of the gallbladder. Eur. J. Surg. Oncol. 28, 4–10 (2002).
2. Wistuba, I. I. & Gazdar, A. F. Gallbladder cancer: lessons from a rare tumour. Nat. Rev. Cancer 4, 695–706 (2004).
3. Hnidal, R. & Shaffer, E. A. Gallbladder cancer: epidemiology and outcome. Clin. Epidemiol. 6, 99–109 (2014).
4. Goetz, T. O. & Paolucci, V. Adequate extent in radical resection of incidental gallbladder carcinoma: analysis of the German Registry. Surg. Endosc. 24, 2156–2164 (2010).
5. Heitz, L., Kratzer, W., Grater, T., Schmidberger, J. & group, E. S. Gallbladder polyps – a follow-up study after 11 years. BMC Gastroenterol. 19, 42 (2019).
6. Inui, K., Yoshino, J. & Miyoshi, H. Diagnosis of gallbladder tumors. Intern. Med. 50, 1133–1136 (2011).
7. Meirelles-Costa, A. L. et al. Are histological alterations observed in the gallbladder precancerous lesions? Clin. (Sao Paulo) 65, 143–150 (2010).
8. Kim, Y. T. et al. Genetic alterations in gallbladder adenoma, dysplasia and carcinoma. Cancer Lett. 169, 59–68 (2001).
9. Feng, Z. et al. The risk factor of gallbladder cancer: hyperplasia of mucous epithelium caused by gallstones associated with p16/CyclinD1/CDK4 pathway. Exp. Mol. Pathol. 91, 569–577 (2011).
10. Kozuka, S., Tsubone, N., Yasui, A. & Hachisuka, K. Relation of adenoma to carcinoma in the gallbladder. Cancer 50, 2226–2234 (1982).
11. Sun, Y., Yang, Z., Lan, X. & Tan, H. Neoplastic polyps in gallbladder: a retrospective study to determine risk factors and treatment strategy for gallbladder polyps. Hepatobiliary Surg. Nutr. 8, 219–227 (2019).
12. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. Cell 61, 759–766 (1990).
13. Lee, S. R., Kim, H. O. & Shin, J. H. Reasonable cholecystectomy of gallbladder polyp - 10 years of experience. Asian J. Surg. 42, 332–337 (2019).
14. Katabi, N. Neoplasia of gallbladder and biliary epithelium. Arch. Pathol. Lab. Med. 139, 1621–1627 (2015).
15. Vicent, S. et al. Experimental models to unravel the molecular pathogenesis, cell of origin and stem cell properties of cholangiocarcinoma. Liver Int. 39, 79–97 (2019).
16. Pai, R. K., Mojtahed, K. & Pai, R. K. Mutations in the RAS/RAF/MAP kinase pathway commonly occur in gallbladder adenomas but are uncommon in gallbladder adenocarcinomas. Appl. Immunohistochem. Mol. Morphol. 19, 133–140 (2011).
17. Itoi, T. et al. APC, K-ras codon 12 mutations and p53 gene expression in carcinoma and adenoma of the gall-bladder suggest two genetic pathways in gall-bladder carcinogenesis. Pathol. Int. 46, 333–340 (1996).
18. Barreto, S. G., Dutt, A. & Chaudhary, A. A genetic model for gallbladder carcinogenesis and its diagnostic potential. Ann. Oncol. 25, 1086–1097 (2014).
19. Roa, I. et al. Preneoplastic lesions and gallbladder cancer: an estimate of the period required for progression. Gastroenterology 111, 232–236 (1996).
20. Akita, M. et al. Intracholangiocytic papillary neoplasms are distinct from papillary gallbladder cancers: a clinicopathological and exome-sequencing study. Am. J. Gastroenterol. 104, 783–791 (2019).
21. Lazcano-Ponce, E. C. et al. Epidemiology and molecular pathology of gallbladder cancer. CA Cancer J. Clin. 51, 349–364 (2001).
22. Ellrott, K. et al. Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. Cell Syst. 6, 271–281 e7 (2018).
23. Alexandrov, L. B. et al. Clock-like mutational processes in human somatic DNA. Nat. Genet. 47, 1402–1407 (2015).
24. Kauto, E. A. et al. Performance evaluation for rapid detection of pan-cancer microsatellite instability with MANTIS. Oncotarget 8, 7452–7463 (2017).
25. Adnan Awad, S. et al. Mutation accumulation in cancer genes relates to nonoptimal outcome in chronic myeloid leukemia. Blood Adv. 4, 546–559 (2020).
26. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. Cell 173, 371–385 e18 (2018).
27. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. Curr. Protoc. Hum. Genet. Chapter 7, Unit7.20 (2013).
28. Reboiuou, S. et al. Genotype-phenotype correlation of CTNNB1 mutations reveals different α-catenin activity associated with liver tumor progression. Hepatology 64, 2047–2061 (2016).
29. Reiter, F. G. et al. Reconstructing metastatic seeding patterns of human cancers. Nat. Commun. 8, 14114 (2017).
30. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol. 35, 1547–1549 (2018).
31. Chang, H. J., Kim, S. W., Kim, Y. T. & Kim, W. H. Loss of heterozygosity in dysplasia and carcinoma of the gallbladder. *Mod. Pathol.* 12, 763–769 (1999).

32. Jain, K. et al. Sequential occurrence of preneoplastic lesions and accumulation of loss of heterozygosity in patients with gallbladder stones suggest causal association with gallbladder cancer. *Ann. Surg.* 260, 1073–1080 (2014).

33. Shen, R. & Seshan, V. E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.* 44, e131 (2016).

34. Roth, A. et al. PyClone: statistical inference of clonal population structure in cancer. *Nat. Methods* 11, 396–398 (2014).

35. Hu, Z., Li, Z., Ma, Z. & Curtis, C. Multi-cancer analysis of clonality and the timing of systemic spread in paired primary tumors and metastases. *Nat. Genet.* 52, 701–708 (2020).

36. Roa, I., de Aretxabala, X., Araya, J. C. & Roa, J. Preneoplastic lesions in hepatocellular carcinoma: synchronous and metachronous. *Pathol. Res. Pract.* 216, 1679–1687 (2019).

37. Yanagisawa, N., Mikami, T., Saegusa, M. & Okayasu, I. More frequent beta-catenin exon 3 mutations in gallbladder adenomas than in carcinomas indicate different lineages. *Cancer Res.* 61, 19–22 (2001).

38. Akki, A. S. et al. Systematic selective sampling of cholecystectomy specimens is adequate to detect incidental gallbladder adenocarcinoma. *Am. J. Surgical Pathol.* 43, 1688–1693 (2019).

39. Bickenbach, K. A. et al. High-grade dysplasia of the cystic duct margin in the absence of malignancy after cholecystectomy. *HPB: Off. J. Int. Hepato Pancreato Biliary Assoc.* 13, 865–868 (2011).

40. Moslim, M. A., Tang, A. & Morris-Stiff, G. Management of high-grade dysplasia of the cystic duct after cholecystectomy. *BMJ Case Rep.* 2017, bcr2016218994 (2017).

41. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760 (2009).

42. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 42, 49–53 (2010).

43. Wang, P. P., Parker, W. T., Branford, S. & Schreiber, A. W. BAM-matcher: a tool for rapid NGs sample matching. *Bioinformatics* 32, 2695–2701 (2016).

44. Saltipante, S. J., Scroggins, S. M., Hampel, H. L., Turner, E. H. & Pritchard, C. C. Microsatellite instability detection by next generation sequencing. *Clin. Chem.* 60, 1192–1199 (2014).

45. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* 17, 1 (2016).

46. Undergauss, A. et al. Primer3—new capabilities and interfaces. *Nucleic Acids Res.* 40, e115 (2012).

47. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput. Biol.* 12, e1004873 (2016).

48. Merdel, 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).

49. Zack, T. I. et al. Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.* 45, 1134–1140 (2013).

50. Boeva, V. et al. Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* 28, 423–425 (2012).

**Acknowledgements**

This study is supported by a faculty scholar award from the University of Texas MD Anderson Cancer Center (to H. Liang), an International Science and Technology Cooperation Project (2016YFE0107100 to H. Zhao), CAMS Innovation Fund for Medical Science (CIFMS, 2017-I2M-4-003 to H. Zhao), National Ten-thousand Talent Program (to H. Zhao), Beijing Science and Technology Cooperation Special Award Subsidy Project (to H. Zhao), and CAMS Initiative for Innovative Medicine (CAMS-2018-I2M-3-001, to H. Zhao). We gratefully acknowledge the support from the Suzhou New District, Jiangsu Province, China (to X.J.). We thank Kamalika Moyumdar for editorial assistance.

**Author contributions**

H. Liang and H.Z. conceived of and designed the research, J.L., K.D., J.L. contributed to the collection of specimens and discussion of clinical significance; X.P. led the data analysis; X.G., H. Li, and X.J. contributed to the data analysis; Y.B., X.Y., D.W., X.L., Y.M., and X.S. contributed to the collection of specimens and performed the experiments. X.P., J.L., H.Z., and H. Liang wrote the manuscript with input from all authors.

**Competing interests**

X.P., X.G., H. Li., and X.J. are full-time employees, and H. Liang is a shareholder and scientific advisor of Precision Scientific Ltd. All other authors declare that they have no competing interests.

**Additional information**

*Supplementary information* The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-25012-9.

**Correspondence** and requests for materials should be addressed to H.Z. or H.L.

**Peer review information** *Nature Communications* thanks Zheng Hu and the other anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permission information** is available at http://www.nature.com/reprints

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.