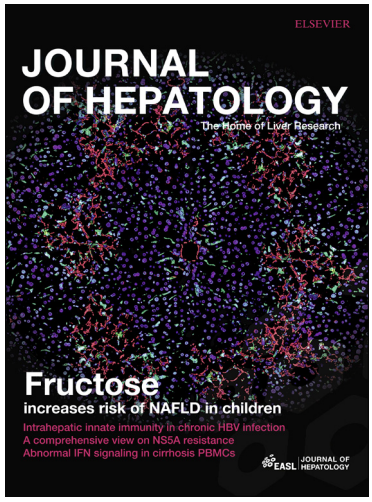
Accepted Manuscript



Genomic Characterization of Biliary Tract Cancers Identifies Driver Genes and Predisposing Mutations

Christopher P. Wardell, Masashi Fujita, Toru Yamada, Michele Simbolo, Matteo Fassan, Rosa Karlic, Paz Polak, Jaegil Kim, Yutaka Hatanaka, Kazuhiro Maejima, Rita T. Lawlor, Yoshitsugu Nakanishi, Tomoko Mitsuhashi, Akihiro Fujimoto, Mayuko Furuta, Andrea Ruzzenente, Simone Conci, Ayako Oosawa, Aya Sasaki-Oku, Kaoru Nakano, Hiroko Tanaka, Yujiro Yamamoto, Kubo Michiaki, Yoshiiku Kawakami, Hiroshi Aikata, Masaki Ueno, Shinya Hayami, Kunihito Gotoh, Shun-ichi Ariizumi, Masakazu Yamamoto, Hiroki Yamaue, Kazuaki Chayama, Satoru Miyano, Gad Getz, Aldo Scarpa, Satoshi Hirano, Toru Nakamura, Hidewaki Nakagawa

PII: S0168-8278(18)30021-7
DOI: https://doi.org/10.1016/j.jhep.2018.01.009
Reference: JHEPAT 6828

To appear in: *Journal of Hepatology*

Received Date: 7 June 2017
Revised Date: 5 January 2018
Accepted Date: 9 January 2018

Please cite this article as: Wardell, C.P., Fujita, M., Yamada, T., Simbolo, M., Fassan, M., Karlic, R., Polak, P., Kim, J., Hatanaka, Y., Maejima, K., Lawlor, R.T., Nakanishi, Y., Mitsuhashi, T., Fujimoto, A., Furuta, M., Ruzzenente, A., Conci, S., Oosawa, A., Sasaki-Oku, A., Nakano, K., Tanaka, H., Yamamoto, Y., Michiaki, K., Kawakami, Y., Aikata, H., Ueno, M., Hayami, S., Gotoh, K., Ariizumi, S-i., Yamamoto, M., Yamaue, H., Chayama, K., Miyano, S., Getz, G., Scarpa, A., Hirano, S., Nakamura, T., Nakagawa, H., Genomic Characterization of Biliary Tract Cancers Identifies Driver Genes and Predisposing Mutations, *Journal of Hepatology* (2018), doi: https://doi.org/10.1016/j.jhep.2018.01.009

Genomic Characterization of Biliary Tract Cancers Identifies Driver Genes and Predisposing Mutations

Christopher P. Wardell[1*], Masashi Fujita[1*], Toru Yamada[2], Michele Simbolo[3], Matteo Fassan[3], Rosa Karlic[4], Paz Polak[5], Jaegil Kim[5], Yutaka Hatanaka[6], Kazuhiro Maejima[1], Rita T. Lawlor[3], Yoshitsugu Nakanishi[2], Tomoko Mitsuhashi[7], Akihiro Fujimoto[1], Mayuko Furuta[1], Andrea Ruzzenente[8], Simone Conci[8], Ayako Oosawa[1], Aya Sasaki-Oku[1], Kaoru Nakano[1], Hiroko Tanaka[9], Yujiro Yamamoto[1,10], Kubo Michiaki[10], Yoshiiku Kawakami[11], Hiroshi Aikata[11], Masaki Ueno[12], Shinya Hayami[12], Kunihito Gotoh[13], Shun-ichi Ariizumi[14], Masakazu Yamamoto[14], Hiroki Yamaue[12], Kazuaki Chayama[11], Satoru Miyano[9], Gad Getz[5], Aldo Scarpa[3], Satoshi Hirano[2], Toru Nakamura[2], and Hidewaki Nakagawa[1,15]

[1]*Laboratory for Genome Sequencing Analysis, RIKEN Center for Integrative Medical Sciences, Tokyo 108-8639, Japan.*

[2]*Department of Gastroenterological Surgery II, Hokkaido University Graduate School of Medicine, Sapporo 060-8638, Japan.*

[3]*Department of Pathology, and Diagnostics, University of Verona, Verona, Italy*

[4]*Bioinformatics Group, Department of Molecular Biology, Division of Biology, Faculty of Science, University of Zagreb, Croatia*

[5]*Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA*

[6]*Research Division of Companion Diagnostics, Hokkaido University Hospital, Sapporo 060-8648, Japan*

[7]*Department of Surgical Pathology, Hokkaido University Hospital, Sapporo 060-8648, Japan.*

[8]*Department of Surgery, General and Hepatobiliary Surgery, University Hospital G.B. Rossi, University and Hospital Trust of Verona, Verona, Italy*

[9]*Laboratory of DNA Information Analysis, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo 108-8639, Japan.*

[10]*Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045, Japan.*

[11]*Department of Medicine & Molecular Science, Hiroshima University School of Medicine, Hiroshima 734-8551, Japan.*

[12]*Second Department of Surgery, Wakayama Medical University, Wakayama 641-8510, Japan*

[13]*Department of Surgery, Osaka Medical Center for Cancer and Cardiovascular Diseases, Osaka 537-8511, Japan.*

[14]*Department of Gastroenterological Surgery, Tokyo Women's Medical University, Tokyo 162-8666, Japan.*
[15]*Correspondence should be addressed to H. Nakagawa (hidewaki@ims.u-tokyo.ac.jp)*
**They contributed equally.*

Electronic word count: 6370 words

Number of figures and tables: 4 figures and 2 tables

Abbreviations. BTC: biliary tract cancer, ICC: intrahepatic cholangiocarcinoma, DCC: distal cholangiocarcinoma, PHC: peri-hilar type of cholangiocarcinoma, GBC: gallbladder cancer, CDC: cystic duct cancer, HCC: hepatocellular carcinoma, WGS: whole genome sequencing, WES: whole exome sequencing, TMA: tissue microarrays, NMF: Nonnegative matrix factorization, OS: overall survival, DFS: disease-free survival, ICGC: International Cancer Genome Consortium, TCGA: The Cancer Genome Atlas, COO: cell-of-origin, VUS: variants of uncertain significance.

Key words: biliary tract cancer, cholangiocarcinoma, genome sequencing, driver gene, cell origin, predisposing mutation

Conflict of interest: No potential conflicts of interest were disclosed by all authors. Financial supports: This work was supported partially by RIKEN President's Fund 2011, Grandin-aid for RIKEN CGM and IMS, JSPS Grants-in-Aid for Scientific Research (KAKENHI) for H.N. (15H04814), the Princess Takamatsu Cancer Research Fund, Italian Ministry of Research (Cancer Genome Project FIRB RBAP10AHJB), Associazione Italiana Ricerca Cancro (AIRC n. 12182), and Fondazione Italiana Malattie Pancreas – Ministero Salute (CUP_J33G13000210001)

Author Contributions: C.P.W., M. F., M.S., M.F., P.P., J.K., R.K., M.F., A.F., G.G., and H.N. performed data analyses. M. S., M.F., K.M., A.O., A.S., K.N., Y.Y., M.K. and H.N. performed sample processing and sequencing. H.T. and S.M. operated the super-computer system. T.Y. and Y.H. performed immunohistochemical analysis. M.S., M.F., R.T.L., Y.N., T.M., Y.H., A.R., S.C., Y.K., H.A., M.U., S.H., K.G., S.A., M.Y., H.Y., K.C., A.S., S.H., and T.N. collected clinical samples and clinical information. C.P.W., M.F., T.M. T.N. and H.N. wrote the manuscript. H.N. conceived the study and led the design of the experiments. H.N. and A.S. contributed to the funding for this study.

ABSTRACT

Background & Aims: Biliary tract cancers (BTCs) are clinically and pathologically heterogeneous with poor response to treatments. Genomic profiling can offer a clearer understanding of their carcinogenesis, classification and treatment strategy. We performed large scale genome sequencing analyses on BTCs to investigate their somatic and germline driver events and characterize their genomic landscape.

Methods: We analyzed 412 BTC samples from Japanese and Italian populations, with 107 whole exome sequencing (WES), 39 whole genome sequencing (WGS), and targeted sequencing of a further 266 samples. The subtypes were 136 intrahepatic cholangiocarcinomas (ICCs), 101 distal cholangiocarcinomas (DCCs), 109 peri-hilar types (PHCs), and 66 gallbladder or cystic duct cancers (GBCs/CDCs). We identified somatic alterations and searched for driver genes in BTCs, and found pathogenic germline variants of cancer-predisposing genes. We predicted cell-of-origin for BTCs by combining somatic mutation patterns and epigenetic features.

Results: We identified 32 significantly and commonly mutated genes including *TP53*, *KRAS*, *SMAD4*, *NF1*, *ARID1A*, *PBRM1*, and *ATR*, some of which negatively affected patient prognosis. A novel deletion of *MUC17* at *7q22.1* affected patient prognosis. Cell-of-origin predictions using WGS and epigenetic features suggest hepatocyte-origin of hepatitis-related ICCs. Deleterious germline mutations of cancer-predisposing genes such as *BRCA1*, *BRCA2*, *RAD51D*, *MLH1*, or *MSH2* were detected in 11% of BTC patients.

Conclusions: BTCs have distinct genetic features including somatic events and germline predisposition. These findings could be useful to establish treatment and diagnostic strategies for BTCs based on genetic information.

Lay Summary: We here analyzed genomic features of 412 BTC samples from Japanese and Italian population. 32 significantly and commonly mutated genes were identified, some of which negatively affected patient prognosis, including a novel deletion of *MUC17* at *7q22.1*. Cell-of-origin predictions using WGS and epigenetic features suggest hepatocyte-origin of hepatitis-related ICCs. Deleterious germline mutations of cancer-predisposing genes were detected in 11% of BTC patients. BTCs have distinct genetic features including somatic events and germline predisposition.

INTRODUCTION

Biliary tract cancer (BTC) or cholangiocarcinoma is a rare cancer worldwide, but prevalent in some areas, where a specific risk factor of environmental exposure is involved in BTC development such as chronic cholangitis[1,2], liver fluke infection in Thailand[1,2], viral hepatitis[1,2], aflatoxin exposure in Chile[3], or other chemical exposure [2,4]. According to its anatomical location, BTCs are mainly classified as intrahepatic (ICC), extra-hepatic bile duct cancer, or gallbladder cancer. The extra-hepatic form is composed of peri-hilar type (PHC or Klatskin tumor) and a distal form (DCC) while gallbladder cancer (GBC) also includes cystic duct carcinoma (CDC). There is some debate about the cellular origin of ICC. BTC cells are presumed to originate from cholangiocytes, but the presence of mixed tumor types in ICC and HCC (hepatocellular carcinoma) with intermediate characteristics between ICC and HCC suggests that a subset of intrahepatic tumors could share a common hepatic progenitor cell origin[5]. Regardless of its location or pathology, BTCs are very aggressive with high metastatic and invasive potential and are difficult to completely resect by surgery due to their anatomical location and spread along the bile ducts. The standard of practice for advanced cholangiocarcinoma is cisplatin or gemcitabine, but the response to these chemotherapy is poor, and consequently they show poor prognosis with only 5-10% five-year survival[1].

Previous studies of the genomic alterations in a variety of BTC types[6–10] have found commonalities such as *TP53*, *KRAS* and *SMAD4* mutations, with a second tier of less frequently mutated genes including *ARID1A*, *CDKN2A*, *IDH1*, *ELF3* and *PIK3CA* that were not seen in all studies. Molecular differences between the subtypes have tended to be in the frequency of mutations in certain genes, rather than different sets of genes being mutated. The genetic features of BTC remain poorly understood and the molecular profiles of BTCs are as heterogeneous as their pathology and biology, making large sample sizes necessary for comprehensive analysis and understanding of its molecular carcinogenesis and clinical associations. To enable this, we performed large-scale genome sequencing analysis of 412 BTC samples from Japanese and Italian populations with various BTC subtypes. We also investigated the genetic heterogeneity of somatic and germline events and cell origin of BTCs and the prognostic value of the observed genetic alterations. These findings indicate that BTCs are quite heterogeneous at the molecular level, but share some distinct genetic features in somatic events and germline predisposition.

MATERIALS AND METHODS

Clinical samples

A series of 218 fresh-frozen tumors and normal tissues from BTC patients in Hokkaido University Hospital and other four Japanese hospitals underwent WES, WGS, and targeted sequencing. They were obtained between 2003 and 2015 and the median follow-up period was 24 months. Their clinico-pathological features are in **Supplementary Table 1**. DNA was extracted from fresh-frozen tumor specimens and adjacent normal tissues or blood. All subjects agreed with informed consent to participate in the study following ICGC guidelines. IRBs at RIKEN (H20-16), Hokkaido University (16-051) and all participating hospitals in this study approved this work.

For Italian samples, a retrospective series (1990-2014) of 194 surgically-resected BTCs were retrieved from the FFPE archives of the ARC-Net Biobank at Verona University Hospital under the local ethics committee approval (n. prog. 1959). All cases were reclassified by two pathologists (MF and AS) according to WHO 2015 as follows: 70 intra-hepatic (ICC), 46 peri-hilar tumor (PHC), 52 distal extra-hepatic tumor (DCC) and 26 gallbladder carcinoma (GBC). Staging was according to AJCC/UICC 7th edition. Matched normal liver was used to determine the somatic nature of mutations.

Library preparation and DNA sequencing

Exome capture was carried out using Nextera Rapid Capture Exomes kits (Illumina). For WGS, DNA was extracted from cancer and normal tissue and 500~600-bp insert libraries were prepared according to the protocol provided by Illumina. The exome-captured or WGS libraries were sequenced on HiSeq2000/2500 with paired reads of 100-125bp. Targeted sequencing for Japanese samples was performed by Illumina HiSeq2000/2500 after capturing by SureSelect XT Custom kit (Agilent Technologies) for 49 candidate driver genes (**Supplementary Table 2**). For Italian samples, targeted sequencing was performed by Ion Proton (Thermo Fisher). Four multigene panels were used: the 50-gene Ion AmpliSeq Cancer Hotspot panel v2 (Life Technologies) and three AmpliSeq custom panel targeting 39 genes not included in the commercial panel. They are listed in **Supplementary Table 2**. 20 ng of DNA were used for each multiplex PCR amplification. The quality of the obtained libraries was evaluated by Agilent2100 Bioanalyzer (Agilent Technologies).

SNV and indel calling

Exome and WGS data were processed as follows. After alignment of reads using BWA[11], PCR duplicates were removed using Picard (http://picard.sourceforge.net/). SNVs were called using MuTect[12]. Indels were called using Strelka[13]. SNVs were further filtered according to the following criteria:

- Minimum depth >= 10x
- Supporting reads >= 3 (at least 1 in either direction)
- Mean mapping quality >= 50
- Mean base quality >= 25
- Minimum variant allele frequency (VAF) >= 0.15
- No strand bias (via Fisher's exact test)
- Sequence must be unique (alignability score of 1 using the GEM mappability program [14])
- 8-oxoguanine (C>A|G>T) oxidation artefacts removed [15]
- Only mutations within the exome capture considered

Targeted sequencing data of Japanese samples were processed in the same way as exome data, except the normal reads from one WGS sample (RK317) was given to MuTect instead of a matched normal. The candidates of these somatic mutations were filtered out using germline variants which were validated by Sanger sequencing for normal tissue DNAs. Targeted sequencing data of Italian samples, including alignment to the hg19 human reference genome and variant calling, was done using the Torrent Suite Software v.4.2 (Thermo Fisher). Filtered variants were annotated using a custom pipeline based on vcflib (https://github.com/ekg/vcflib), SnpSift [16], the Variant Effect Predictor (VEP) software [17] and the NCBI RefSeq database.

Significantly mutated genes

Significantly mutated genes were identified using MutSigCV [18]. This analysis was performed once for the 146 WES and WGS samples and once for the complete set of 412 samples covering a subset of the exome. Genes with a minimum of 3 missense mutations at the same position were tested for activation bias using Fisher's exact test comparing the number of activating and inactivating mutations. Similarly, genes with at least 3 inactivating mutations (small insertions and deletions and nonsense SNVs or those impairing stop or start codons and splice sites) were tested for inactivation bias.

Mutational signatures and cell-of-origin analysis

The R package NMF[19] was used to perform non-negative matrix factorization on counts of the 3-base mutation and context types in all WES and WGS samples. Similarity to existing COSMIC signatures was determined using cosine similarity. Cell-of-origin (COO) analysis was performed as previously described [20]. Briefly, we compared the genomic distribution of somatic mutations to 424 epigenetic features that were measured by the Epigenome Roadmap consortium, which were derived from 106 different cell types from 45 different tissue types and comprised eight different types of variables, including DNase I hypersensitivity and various histone modifications. The comparison of individual epigenomic features with local mutation density revealed that the genomic distribution of chromatin features corresponding to the tumor's cell type of origin is more strongly associated with local mutation density than the distribution of features found in unrelated cell types. For each cancer type we counted the overall number of mutations in all individual cancer genomes belonging to that cancer type, and determined the mutation densities for all possible types of mutations in each cancer types by counting different types of mutations in 1-Mb windows and normalizing for the sequence composition of each window. For each individual cancer genome, we predicted the density of mutations using Random Forest regression with tenfold cross-validation. We used the full set of features and determined the top 20 features according to the variable importance measure, and calculated the enrichment of each tissue type among the top 20 features using the hypergeometric test and chose the tissue showing the most significant enrichment as the most likely tissue of origin for the individual cancer genome [20].

DNA arrays and copy number analysis

HumanOmniExpressExome (Illumina) arrays were performed on 95 sets of cancer and noncancer DNAs and their copy number alterations were analyzed using GISTIC2 [21]. Three arrays were excluded after segmentation using PSCBS R package [22] due to quality issues.

Germline variant calling

Germline SNVs and indels were called using Genome Analysis Toolkit version 3.6 (http://software.broadinstitute.org/gatk/). Among the exome and WGS BAM files used for somatic variant calling, those of normal tissues were subjected to base quality score recalibration and HaplotypeCaller. The called variants were filtered by applying variant quality score recalibration, retaining either splicing variants or exonic variants that are not synonymous, and discarding common

variants (allele frequency >1%). Allele frequencies in the Japanese population were obtained from the human genetic variation database in Kyoto University [23] and the integrative Japanese genome variation database in Tohoku Medical Megabank [24]. Allele frequencies in the 1000 Genome Project were also used. The variants in **Table 2** are splicing, stop-gain, frameshift, or (likely) pathogenic ones according to the ClinVar database.

FGFR fusion detection by RNA sequencing

To detect FGFR fusion transcripts, we extracted RNAs from the ICC samples, and RNA-Seq or targeted RNA-Seq was carried out for 63 ICC samples for which high-quality RNA was available. The high-quality RNA was subject to polyA+ selection and chemical fragmentation, and the 100-200 base RNA fraction was used to construct cDNA libraries according to Illumina's protocol. Alternatively, we generated RNA-Seq libraries after capturing the known fusion transcripts (507 genes) including FGFR1/2 according to the protocol of TruSight RNA Fusion Panel (Illumina). Sequencing was performed on HiSeq2500 using the standard paired-end 125bp protocol. RNA-Seq reads were adaptor-trimmed using Cutadapt and mapped to GRCh37 using STAR. Gene fusions were detected using fusionfusion (https://github.com/Genomon-Project/fusionfusion).

Construction of tissue microarray (TMA) and immunohistochemical analysis

A total of 121 surgical specimens of DCCs and PHCs obtained between 1995 and 2006 at the Department of Gastroenterological Surgery II, Hokkaido University Graduate School of Medicine, were used in this study. The median follow-up period was 26 months (range 0–151 months), and 89 patients (76.1%) died during follow-up. None of the patients received chemotherapy preoperatively. Their clinical features are in **Supplementary Table 1**. All informed consent processes for this study were conducted in accordance with the guidelines of the Hokkaido University Institutional Review Board. TMA blocks were constructed using a manual tissue microarrayer (JF-4; Sakura Finetek Japan, Tokyo, Japan) with a 2.0-mm diameter needle from representative tumor areas. The array blocks were sliced into 4-μm-thick serial sections and mounted on glass slides. To confirm the histological diagnosis and adequacy of tissue sampling, TMA sections stained with hematoxylin and eosin (H&E) were examined by an experienced pathologist. TMA sections were deparaffinized in xylene and rehydrated through a graded ethanol series. Heat-induced antigen retrieval was carried out in high-pH antigen retrieval buffer (Dako, Glostrup, Denmark). Endogenous peroxidase was blocked by incubation in 3% H_2O_2 for 5 min. The primary antibody against MUC17 (polyclonal [HPA031634, 1:100]; Atlas Antibodies, Bromma,

Sweden) was applied for 30 minutes. These sections were visualized by the HRP-labeled polymer method (EnVision FLEX system in human tissues; Dako). Immunostained sections were counterstained with hematoxylin, dehydrated in ethanol, and cleared in xylene. In TMA analysis, the immunohistochemistry for MUC17 were evaluated using the H-score system [25] by two researchers who were blinded to the patients' clinical information. Samples were considered high- expression or low-expression on the basis of the median cut-off, or cut-off value determined by ROC.

Survival analysis

Survival analysis was performed using the Surv package in R. Univariate survival was performed using Cox regression and the log rank test. Multivariate Cox regression used stepwise selection for covariate selection, where covariates were required to have a significance below 0.1 to enter the model and below 0.05 to remain in the model.

RESULTS

Mutation frequency and significantly mutated genes. The analysis workflow is shown in **Supplementary Figure 1**. 107 WES tumors and matched controls were sequenced to mean depths of 148-fold and 95-fold respectively, while 39 WGS tumors and matched controls were sequenced to mean depths of 38-fold and 26-fold (**Supplementary Table 3**). We detected 11,543 SNVs and 140 indels after stringent filtering. In the exome, samples contained a mean of 29 SNVs and 0.5 indels (4,125 SNVs and 70 indels total, 0.81 mutations/Mb), with the exception of 6 samples (4% of WES/WGS samples: HK08, HK15, HK67, HK101, RK308, and RK360) which were classified as hypermutated. These contained a mean of 1,236 SNVs and 12 indels (7,418 SNVs and 70 indels total, 33.83 mutations/Mb). A full list of all SNVs and indels is in **Supplementary Table 4**. Significantly mutated genes in the WES and WGS samples were identified using MutSigCV. *TP53* (24 samples, 16.3%) and *KRAS* (8 samples, 5.4%) were found to be significant ($q<0.1$) in the WES and WGS samples (**Supplementary Table 5**). To search for additional driver genes in BTC and its subtypes, we performed targeted deep sequencing on 49 candidate genes for an additional 72 Japanese BTCs and on 89 genes for 194 Italian BTCs (**Supplementary Tables 2**) and combined these data (**Supplementary Figure 1**). Analysis on the entire series of 413 BTCs found 30 significantly mutated genes (**Table 1** and **Supplementary Figure 2**). In addition, *KMT2C* (also known as *MLL3*) and *KMT2D* (also known as *MLL2*) were mutated in more than 5% of the 413 samples (**Supplementary Table 6**), which are also

driver candidates in BTCs. RNA-seq analysis detected *FGFR2* fusion transcripts in four cases among the 63 RNA-available Japanese ICCs (**Supplementary Figure 3**). Summary of the significantly and frequently mutated genes in BTCs are shown in **Figure 1**.

Pathway analysis and subtype enriched features. 268 samples (64.9%) had a mutation in at least one of the 32 genes listed in **Figure 1**, which can be subdivided into three broad functional categories of increasing size; DNA maintenance genes, epigenetic genes and signaling pathways (**Supplementary Figure 4**). Subtype-specific biases were present in a number of genes and functional groups. ICC samples typically contained more mutated epigenetic genes, whereas extrahepatic subtypes contained more mutated cell cycle genes. ICC was significantly enriched for mutations in the chromatin remodeling gene *BAP1* and the DNA methylation gene *IDH1*, which with one exception were in mutually exclusive samples ($p<0.005$, see **Supplementary Figure 4**). *IDH1* mutants were also mutually exclusive from *TET1* mutants (**Figure 1**), which is interesting given that *IDH1* mutants have been shown to correlate with high *TET1* expression in glioma[8]. Gallbladder subtypes were significantly enriched for *TP53* mutations and extrahepatic subtypes were significantly enriched for mutations in *TP53, KRAS, SMAD4* and *ERBB3 (*$p<0.005$*). KRAS* showed strong activation bias, with the majority of samples containing missense mutations affecting the G12 residue, a known oncogenic hotspot. *IDH1* had a single site affected in 13/14 samples and multiple hotspots were found in *PIK3CA*. Six samples with activating mutations in *CTNNB1* were also identified. This gene is activated by nonsynonymous mutations that fall in a 14 amino acid window toward the N-terminus of the protein. Activation and inactivation bias are summarized in **Supplementary Table 7** and **8**, respectively.

Copy number analysis. Copy number differences are common aberrations, and we performed SNP array analysis on 95 samples among WES samples using GISTIC2. Broadly, chromosome *1q* and *19q* were found to be frequently amplified and *4p, 4q, 9p* and *14q* frequently deleted (**Supplementary Table 9**). GISTIC2 detected focal changes in copy-numbers (**Supplementary Figure 5** and **Supplementary Table 10**). A 350-kb region at *17q12* was amplified in 30/92 (32.6%) samples ($p=4.61\times10^{-7}$), which contains the cytokine genes *CCL3L3* and *CCL4L2*, as well as four paralogs of the known oncogene *TBC1D3*, which is amplified in 15% of primary prostate tumors [26]. The tumor suppressor gene *CDKN2A* at *9p21.3* was deleted in 46/92 (50%) samples ($q=2.61\times10^{-22}$). Biallelic

inactivation of *CDKN2A* appeared to be relatively rare [9], with only one sample containing a nonsynonymous *CDKN2A* mutation and one sample showing a low signal value enough to be a homozygous deletion. The most frequently altered region was a 100-kb region in *7q22.1* containing the mucin genes *MUC12* and *MUC17* which was deleted in 59/92 (64.1%) samples ($q=2.29x10^{-13}$). Notably, the only focal copy number change to carry a survival effect was this *7q22.1* deletion. Patients with this deletion suffered a median reduction in progression free survival of 644 days ($p=8.19x10^{-4}$) and overall survival of 428 days (p=0.0133) (**Figure 2**). Patients with lymph node metastases were more likely to have this *7q22.1* deletion (p=0.014, 81% sensitivity and 40% specificity). The two candidate genes at this *7q22.1* region are the mucins *MUC12* and *MUC17* (**Supplementary Figure 5)**. Quantitative PCR analysis of the 92 BTC samples showed that *MUC17* expression was significantly decreased in samples with the deletion ($p=5.3x10^{-3}$), but *MUC12* expression was barely detectable in either case (**Supplementary Figure 6**), suggesting that *MUC17* is the gene targeted in this recurrent deletion event and may drive the survival differences seen in these samples. The identified survival trend did not appear to be shared by the 9 samples with missense mutations in *MUC17,* suggesting that complete ablation of this gene is required for the survival effect to occur. BTCs showing *7q22.1* deletion *(MUC17* deletion) were enriched in the samples that had no mutations of the known driver genes detected (**Figure 1**), indicating that this deletion might be one of the driver genes in BTCs.

To validate the clinical effect of MUC17 expression, we performed immunohistochemical analysis on TMA where 121 specimens of DCC and PHC were spotted. Immunohistochemistry and calculation of H-scores [25] indicated lower or loss of MUC17 expression in the cell membrane or the cytoplasm of cancer cells (**Supplementary Figure 7**). We observed a trend that BTCs with lower MUC17 expression showed worse survival (**Supplementary Figure 4**). There is a trend that loss of MUC17 expression was associated with portal vein invasion or micro-vessel invasion (p-value<0.05, **Supplementary Table 11**).

Survival effects of mutations and clinical factors. To clarify the clinical significance of these driver candidates, we examined the association of these mutations and prognosis of 412 BTC cases with both univariate and multivariate analysis of overall survival (OS) and disease-free survival (DFS), shown in **Figure 2** and **Supplementary Table 12**. In univariate analysis, the deletion of *7q22.1* had a significantly negative impact on both DFS (n=52, $p= 3x10^{-4}$) and OS (n-47, p=0.010). Strong negative

effects on OS were observed in patients harboring mutations in *ARID1A* (n=22, p=0.0011) and *KRAS* (n=63, p=0.0042). Smokers and ex-smokers had significantly worse prognosis, both in DFS (n=67, $p=3x10^{-3}$) and OS (n=57, $p=4x10^{-3}$). Multivariate modelling showed TNM and smoking status to be independent variables **Supplementary Table 12**.

Mutational signature analysis. The type and context of somatic mutations can be used to generate distinct mutational signatures that may point at the etiology of cancers [27]. Nonnegative matrix factorization (NMF) extracted three mutational signatures in WES and WGS data of the non-hypermutated cases. These signatures were compared to the 30 COSMIC signatures, revealing close matches to COSMIC Signatures 1 (5-methylcytosine deamination, linked to ageing), Signature 2 (AID/APOBEC deaminases) and Signature 5 (related with age or nucleotide excision repair deficiency) [28] with 0.82, 0.76 and 0.91 cosine similarity, respectively. **Figure 3** shows the samples clustered according to the proportional contribution of each signature per sample. The signature activity data can be divided into three signature clusters, with one signature dominating each. All signatures occurred in varying proportions, with each signature contributing the majority of mutations in multiple samples. Overall, signature 1 was the most broadly occurring (40% total contribution), with the remainder split between signatures 2 (31%) and 5 (29%). Samples in the cluster with a high exposure to Signature 5 were significantly more likely to be ICCs infected with HBV or HCV (p=0.004), suggesting a link between this signature and hepatitis viral infection. An identical analysis was performed using an external data set of 103 BTC samples [9] which produced very similar results (**Supplementary Figure 8**).

Cell-of-origin analysis. The cell-of-origin (COO) of a cancer can be determined by comparing the genomic distribution of mutations to the chromatin organization of specific cell types [20], as mutations are more likely to occur in open, transcriptionally active chromatin. Positively identifying the COO of a cancer could help classify disease subtypes for further study and treatment. We predicted the COO of our BTC subtypes and three hepatocellular carcinoma (HCC) series from the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA). 86 of the 97 (88.7%) HCC samples were classified as originating from hepatocytes (**Figure 4a**). Conversely, the most common classifications of the 39 WGS BTC samples were liver (13/39 33.3%), or epithelial cell types (14/39

35.9%). Proportionally, it was more common for ICC samples to be classified as originating from hepatocytes than epithelia (43.5% and 17.4% respectively, **Figure 4b**), whereas the reverse was true for DCC samples (0% and 83.3% respectively). Using the CAGE dataset in FANTOM5 Consortium [29], we validated that epigenetic features of DCC and GBC/CDC were matched with those of gallbladder (**Supplementary Figure 9**), although epigenetic features of some ICCs were still matched with those of hepatocytes. Liver-COO tumors were enriched with ICCs with hepatitis background (viral infection or fibrosis F3/F4, **Figure 4C**) and showed significantly more Signature 5 than non-liver-COO ICC samples ($p = 0.04$ by the Wilcoxon rank sum test, **Supplementary Figure 10**).

Germline variants in cancer-predisposing genes in BTC patients. Germline DNA-mismatch repair deficiency (Lynch syndrome) is related with BTC [30], and carriers of germline mutations in *BRCA2* are at high risk of BTC as well as pancreatic cancer [31]. We analyzed germline variants of selected cancer-predisposing genes, including *BRCA1, BRCA2, RAD51D, MSH2, MLH1, MSH6,* and *TP53,* and annotated their significance. As a result, 8 BTC patients were found to have deleterious germline mutations of *BRCA1, BRCA2,* or *RAD51D,* and 6 BTC patients had deleterious germline mutations of *MLH1, MSH2, POLD1,* or *POLE* (**Table 2**). Two BTC patients had deleterious germline mutations of *TP53* or *ATM*. We verified these germline variants by Sanger sequencing. In total, 11% (16/147) of Japanese BTC patients had deleterious germline mutations of the cancer-predisposing genes. Two (HK15 and HK08) out of the six hypermutated BTC cases were found to have deleterious germline mutations of *MLH1* or *MSH2* and developed gastric and colon cancers (**Table 2**), indicating typical Lynch syndrome. An additional 80 variants of uncertain significance (VUS) in these cancer-predisposing genes, some of which may be pathogenic, were found in 75/147 (51.0%) Japanese BTC patients (**Supplementary Table 13**).

DISCUSSION

This analysis of 412 BTC samples in two populations represents the largest series studied to date. We identified several novel features which are enriched in certain subtypes and others with impacts on both OS and DFS. This study identified a variety of targets of biological and therapeutic relevance, some of which vary by disease subtype. While larger datasets are always beneficial, it appears that the inherent heterogeneity of BTC is beginning to resolve, although 31 samples (7.5%) did not harbor any of the

copy number alterations or mutated genes identified, demonstrating that gaps in our knowledge remain and other genetic or epigenetic alterations can drive BTC development.

Of the 32 frequently mutated genes identified in **Table 1**, 21 are actionable according to the TARGET database by the Broad Institute (http://archive.broadinstitute.org/cancer/cga/target). 241 samples (58.4%) had one or more actionable lesions and 123 samples (29.8%) had 2 or more actionable lesions, representing broad opportunities for treatment. Although the number of hypermutated samples was too small to identify which specific genes were driving the phenotype, all of the hypermutated samples harbored somatic or germline mutations in at least one known DNA mismatch repair gene or DNA polymerase (*MLH1, MSH2, MSH6, POLE, or POLQ*). Hypermutated cancers may be excellent candidates for immunotherapy and immune checkpoint inhibitors [32] because their high immunogenecity by neo-antigen presentation. There is evidence that ICC is more frequently an epigenetic disease [5, 6], whereas extrahepatic and gallbladder subtypes are driven by mutations in *TP53* and cell cycle genes [7]. For example, the enrichment and mutual exclusivity of *IDH1* and *BAP1* mutations in ICC, an occurrence that has been previously observed [6,33] which provides an unknown fitness advantage. BAP1 is a deubiquitinase involved in chromatin modification and DNA damage response [33] whereas dysregulation of IDH1 may cause global DNA methylation dysfunction [34, 35]. Several mutated genes were found to have negative survival effects, and one of the strongest survival effects belonged to the novel recurrent deletion at *7q22.1* which excises *MUC17*. *MUC17* is a transmembrane glycoprotein of largely unknown function, which has been found to be part of the glycocalyx of enterocytes [36] and may have an immune signaling function. *MUC17* was previously found to be significantly mutated in glioblastoma multiforme [37] and mucins in general are beginning to be broadly recognized for their roles in cancer. However, details of its function in BTC are still unclear and should be investigated by further analysis.

Our new findings suggest a link between cell-of-origin and BTC subtypes, specifically that some ICCs may develop from hepatocytes or their progenitor cells. Pluripotent progenitor cells in the liver can differentiate into hepatocytes and cholangiocytes during liver regeneration [38]. It has also been shown that some ICCs originate from either pluripotent progenitor cells or multiple cell types [39, 40], and our COO analyses using whole genome mutational features and the tissue-specific epigenetic features demonstrated that ICCs that were predicted to be of hepatocyte-origin were associated with virus infection or chronic hepatitis [5,40].

Notably, we found at least 11% of BTC cases had deleterious germline mutations in cancer-predisposing genes. BTC is a tumor related with Lynch syndrome, which has germline mutations of DNA mismatch repair genes [28], and BTC also occurs in carriers of *BRCA* mutations [31]. Universal tumor screening [41] for these cancer-predisposing genes in general BTC cases might be beneficial to BTC patients and their family members to assess their cancer development risk and the effectiveness of new anti-cancer drugs such as immune checkpoint inhibitors and PARP inhibitors.

ACKNOWLEDGEMENTS

The super-computing resource 'SHIROKANE' was provided by Human Genome Center, The University of Tokyo.

Databases

- TARGET; http://archive.broadinstitute.org/cancer/cga/target
- ClinVar database; http://www.ncbi.nlm.nih.gov/clinvar/.
- Human genetic variation database in Kyoto University; http://www.hgvd.genome.med.kyoto-u.ac.jp/
- Integrative Japanese Genome Variation Database; https://ijgvd.megabank.tohoku.ac.jp/
- FANTOM5; http://fantom.gsc.riken.jp/5/data/

Data deposition

We deposited the raw sequence data in JGA.
Submission: JGA00000000119
Study: JGAS00000000109
Dataset: JGAD00000000117-JGAD00000000118

References

1. Malhi, H. and Gores, G.J. Cholangiocarcinoma: modern advances in understanding a deadly old disease. *J Hepatol* **45**:856-867 (2006).

2. Palmer, W.C. and Patel, T. Are common factors involved in the pathogenesis of primary liver cancers? A meta-analysis of risk factors for intrahepatic cholangiocarcinoma. *J Hepatol* **57**:69-76 (2012).

3. Nogueira, L. *et al.* Association of aflatoxin with gallbladder cancer in Chile. *JAMA* **313,** 2075–7 (2015).

4. Yamada, K., Kumagai, S., Nagoya, T. & Endo, G. Chemical exposure levels in printing workers with cholangiocarcinoma. *J. Occup. Health* **56,** 332–8 (2014).

5. Raggi C, Invernizzi P, and Andersen JB. Impact of microenvironment and stem-like plasticity in cholangiocarcinoma: molecular networks and biological concepts. *J Hepatol* 62:198-207 (2015).

6. Simbolo, M. *et al.* Multigene mutational profiling of cholangiocarcinomas identifies actionable molecular subgroups. *Oncotarget* **5,** 2839–52 (2014).

7. Li, M. *et al.* Whole-exome and targeted gene sequencing of gallbladder carcinoma identifies recurrent mutations in the ErbB pathway. *Nat. Genet.* **46,** 1–7 (2014).

8. Zou, S. *et al.* Mutational landscape of intrahepatic cholangiocarcinoma. *Nat. Commun.* **5**, 5696 (2014).

9. Nakamura, H. *et al.* Genomic spectra of biliary tract cancer. *Nat. Genet.* **47,** 1003–10 (2015).

10. Farshidfar, F. *et al.* Integrative genomic analysis of cholangiocarcinoma identifies distinct IDH-mutant molecular profiles. *Cell Rep*. **18**, 2780-94 (2017)

11. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–60 (2009).

12. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31,** 213–9 (2013).

13. Saunders, C. T. *et al.* Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28,** 1811–1817 (2012).

14. Derrien, T. *et al.* Fast computation and applications of genome mappability. *PLoS One* **7,** e30377 (2012).

15. Costello, M. *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* **41,** e67 (2013).

16. Cingolani, P. *et al.* Using Drosophila melanogaster as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front. Genet.* **3,** (2012).

17. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26,** 2069–70 (2010).

18. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer associated genes. *Nature* **499,** 214–8 (2013).

19. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11,** 367 (2010).

20. Polak, P., Karlic, R., *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518,** 360–364 (2015).

21. 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).

22. Bengtsson, H., Neuvial, P. & Speed, T. P. TumorBoost: normalization of allele-specific tumor copy numbers from a single pair of tumor-normal genotyping microarrays. *BMC Bioinformatics* **11,** 245 (2010).

23. Higasa, K. *et al.* Human genetic variation database, a reference database of genetic variations in the Japanese population. *J Hum Genet* **61**:547-553 (2016).

24. Yamaguchi-Kabata, Y. *et al.* iJGVD: an integrative Japanese genome variation database based on whole-genome sequencing. *Human Genome Variation* **2**, 15050 (2015).

25. Hirsch, F.R., *et al.* Epidermal growth factor receptor in non-small-cell lung carcinomas: Correlation between gene copy number and protein expression and impact on prognosis. *J Clin Oncol* **21**:3798-3807 (2003).

26. Hodzic, D. *et al.* TBC1D3, a hominoid oncoprotein, is encoded by a cluster of paralogues located on chromosome 17q12. *Genomics* **88,** 731–6 (2006).

27. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500,** 415–421 (2013).

28. Alexandrov, L. B. *et al.* Mutational signatures associated with tobacco smoking in human cancer. *Science* **354**, 618-622 (2016).

29. Forrest, A.R., *et al.* A promoter level mammalian expression atlas. *Nature* **507**, 462-470 (2014).

30. Mecklin, J.P., Jarvinen, H.J., and Virolainen, M. The association between cholangiocarcinoma and hereditary nonpolyposis colorectal carcinoma. *Cancer* **69**:1112-4 (1992).

31. The Breast Cancer Linkage Consortium. Cancer risks in BRCA2 mutation carriers. *J Nat Cancer Inst* **91**:1310-16 (1999).

32. Angelova, M. *et al.* Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. *Genome Biol.* **16,** 64 (2015).

33. Jiao, Y. *et al.* Exome sequencing identifies frequent inactivating mutations in BAP1, ARID1A and PBRM1 in intrahepatic cholangiocarcinomas. *Nat. Genet.* **45,** 1470–1473 (2013).

34. Turkalp, Z., Karamchandani, J. & Das, S. IDH mutation in glioma: new insights and promises for the future. *JAMA Neurol.* **71,** 1319–25 (2014).

35. Müller, T. *et al.* Nuclear exclusion of TET1 is associated with loss of 5hydroxymethylcytosine in IDH1 wild-type gliomas. *Am. J. Pathol.* **181,** 675–83 (2012).

36. Pelaseyed, T. *et al.* The mucus and mucins of the goblet cells and enterocytes provide the first defense line of the gastrointestinal tract and interact with the immune system. *Immunol. Rev.* **260,** 8–20 (2014).

37. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505,** 495–501 (2014).

38. Duncan, A. W., Dorrell, C. & Grompe, M. Stem cells and liver regeneration. *Gastroenterology* **137,** 466–81 (2009).

39. Fan, B. *et al.* Cholangiocarcinomas can originate from hepatocytes in mice. *J. Clin. Invest.* **122,** 2911–5 (2012).

40. Fujimoto, A. *et al*. Whole-genome mutational landscape of liver cancers displaying biliary phenotype reveals hepatitis impact and molecular diversity. *Nat. Commun*. **6**, 6120 (2015).

41. Moreira, L. *et al*. Identification of Lynch syndrome among patients with colorectal cancer. *JAMA* **308**:1555-65 (2012).

Figure legends

Figure 1

Significantly and frequently mutated genes in BTCs. Each column represents one of the 412 BTC samples and each row represents a feature. From top to bottom, the subtype and data set of each sample is given, the amplification of the *17q12* region containing *TBC1D3B*, the deletions of the *9p21.3* regions containing *CDKN2A* and the *7q22.1* regions containing *MUC17*, the presence of *FGFR2* fusion, the viral infection status of each patient, their gender and the presence and type of mutations in each of the 32 genes identified as either significantly mutated ($q<0.1$ in MutSigCV) or frequently mutated (minimum 5% of samples). A color key is at the top right. The bar plot on the right shows the number of samples containing a mutation in each gene.

Figure 2

Impact of specific mutations on overall survival (OS) and disease-free survival (DFS) in BTCs. (a) Negative effects on OS and DFS of the frequently deleted region containing *MUC17* (*7q22.1*) in Japanese BTC samples. (b) *ARID1A* and *KRAS* mutations associated with negative effects on OS in the combined BTC dataset. (c) Negative effects of smoking status of the BTC patients on both their OS and DFS.

Figure 3

Mutational signatures detected in BTC. (a) Proportional barplot showing relative exposure to each of the 3 signatures. Bars colors match the labels given in part b and the dendrogram is colored to show 3 clusters, each dominated by a single signature. The subtype, viral infection status, predicted cellular

origin and smoking status of each sample is encoded at the top of the barplot. (b) The 3 mutational signatures detected in the BTC samples, annotated with the corresponding COSMIC signature determined by cosine similarity. Signature A (COSMIC Signature 5, cosine similarity = 0.90), Signature B (COSMIC Signature 2, cosine similarity = 0.76), and Signature C (COSMIC Signature 1, cosine similarity = 0.82).

Figure 4

Cell-of-origin analysis on BTCs and liver cancers. (a) Results for four WGS series; the BTC samples from this WGS study (n=39) and three ICGC liver cancer WGS data sets (totally n=97). HCC samples were overwhelmingly predicted to originate from liver cells (*), whereas BTC samples were split between liver origin and mostly epithelial cell types. (b) Proportional results for each BTC subtype. DCC samples were mostly predicted to originate from other epithelial cell types and never from liver cells. Conversely, ICC samples were frequently predicted to originate from liver cells (*). (c) Higher proportion of ICC with hepatitis background (N=13, virus infection or liver fibrosis F3/F4 according to the Inuyama classification) was predicted to be liver-COO, while the liver-COO proportion of ICC without hepatitis background (N=10) was similar to that of other subtypes of BTCs (N=16).

Table 1 Frequently mutated genes in BTCs

GENE	MUTATION# IN 147 BTC	FREQUENCY IN 147 BTC	MUTATION# IN 412 BTC	FREQUENCY IN 412 BTC	MUTSIGCV P-VALUE	MUTSIGCV Q-VALUE
TP53*	24	16%	106	26%	0	0
KRAS*	8	5%	70	17%	0	0
SMAD4	5	3%	34	8%	0	0
NF1	9	6%	25	6%	2.01E-07	0.000147
ARID1A	3	2%	25	6%	4.1E-14	3.74E-11
PBRM1	3	2%	24	6%	1.75E-13	1.52E-10
KMT2D†	6	4%	24	6%	0.003226	1
ATR	8	5%	23	6%	2.7E-06	0.001758
PIK3CA	6	4%	22	5%	0	0
ERBB3	8	5%	19	5%	1.49E-09	1.24E-06
KMT2C†	6	4%	19	5%	0.000236	0.138986
PIK3C2G	2	1%	18	4%	0	0
APC	6	4%	18	4%	1.85E-07	0.000141
BAP1	1	1%	17	4%	0	0
POLQ	6	4%	16	4%	6.08E-07	0.000427
ARID2	7	5%	15	4%	3.37E-08	2.67E-05
IDH1	1	1%	14	3%	0	0
TET1	4	3%	13	3%	3.62E-05	0.022033
CTNNB1	3	2%	12	3%	0	0
BRAF	3	2%	12	3%	0	0
TGFBR2	1	1%	11	3%	0	0
PTEN	1	1%	11	3%	0	0
DNMT3A	5	3%	11	3%	0	0
FBXW7	2	1%	10	2%	0	0
ELF3	5	3%	10	2%	0	0
CDKN2A	3	2%	10	2%	0	0
MSH6	4	3%	8	2%	1.15E-05	0.007271
STK11	3	2%	6	1%	0	0
RNF43	3	2%	6	1%	0	0
NRAS	0	0%	6	1%	1.14E-06	0.000773
MLH1	1	1%	6	1%	0	0
TGFBR1	3	2%	5	1%	0	0

* denotes significance in 147 WES and WGS samples (2 genes). † denotes genes present in >5% of 412 samples, although not statistically significant (2 genes, excluding *TTN*). All non-marked genes were significant in 412 samples (30 genes).

Table 2 Deleterious germline mutations in the cancer-predisposing genes in 146 Japanese BTC patients

ID	Age/Sex	Subtype	Mutation position	Gene	Nucleotide change	Protein change	Significance in Clinvar (20150629)	Japanese allele frequency		Other malignacy	Cancer family history
								HGVD_V2[1)]	ToMMo_2KJP[2)]		
HK09	58M	ICC	chr17:41,243,813	*BRCA1*	c.3593_3594insATAG	p.S1198fs	.	.	.	(-)	(-)
HK12	72M	ICC	chr17:41,258,497	*BRCA1*	c.47T>A	p.L16X	Pathogenic	.	0.0002	(-)	(-)
HK31	65F	KCC	chr13:32,930,687	*BRCA2*	c.7558C>T	p.R2520X	Pathogenic	.	.	(-)	(-)
HK58	74M	ECC	chr13:32,914,066-3,2914,069	*BRCA2*	c.5574_5577delAATT	p.T1858fs	Pathogenic	0.002326	.	**gastric cancer, pharyngeal cancer**	(-)
RK303	76M	ICC	chr13:32,937,642	*BRCA2*	c.8303T>C	p.L2768P	Pathogenic			(-)	(-)
RK309	56M	ICC	chr13:32,954,022	*BRCA2*	c.9090insA	p.T3030fs	Pathogenic			(-)	gastric cancer, lung cancer
HK119	75M	ECC	chr17:33,443,903	*RAD51D*	c.298C>T	p.R100X	.	.	.	(-)	(-)
HK61	73M	ECC	chr17:33,428,057	*RAD51D*	c.964-2 A>T	splicing alteration	.	0.000907	0.0005	(-)	(-)
HK113	64M	KCC	chr3:37,090,506	*MLH1*	c.1378C>A	p.Q460K	Pathogenic	0.001093	0.0002	(-)	(-)
HK15	71M	ICC	chr3:37,038,115	*MLH1*	c.122A>G	p.D41G	Pathogenic	.	.	**colon cancer, rectal cancer, duodenal cancer**	(-)
RK422	70M	ECC	chr3:37,067,242	*MLH1*	c.430C>T	p.R144C	Likely pathogenic	0.001361	0.0012	liver cancer	(-)
HK08	50M	ICC	chr2:47,643,569	*MSH2*	c.1076+1 G>A	splicing alteration	Pathogenic	.	.	**gastric cancer, colon cancer**	(-)
HK52	81M	ECC	chr19:50,910,674	*POLD1*	c.1775+1 T>C	splicing alteration	.	.	.	(-)	(-)
HK30	52M	KCC	chr12:133,253,971	*POLE*	c.779delG	p.R260fs	.	.	.	(-)	(-)
RK361	63M	GBC	chr17:7,578,406	*TP53*	c.128G>A	p.R43H	Pathogenic		.	**gastric cancer**	lung cancer
HK110	60F	ECC	chr11:108,203,578-108,203,582	*ATM*	c.7878_7882delTTATA	p.A2626fs	.	.	.	(-)	(-)

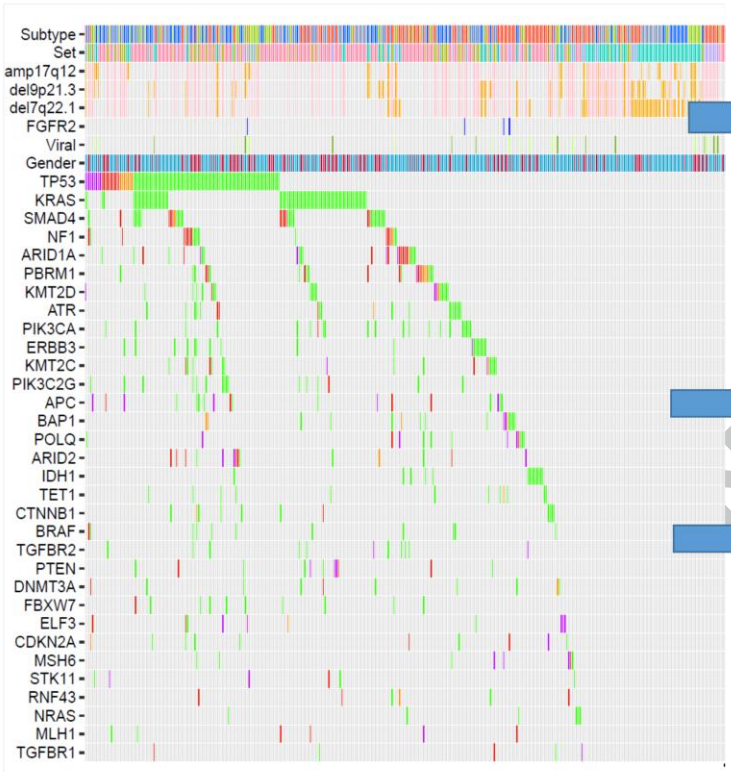
[1)] Human genetic variation database provided by Kyoto University from 3,248 Japanese individuals, [2)] Integrative Japanese Genome Variation Database provided by ToMMo (Tohoku Medical Megabank Organization) from 2,049 Japanese whole genome sequencing.

HIGHLIGHT

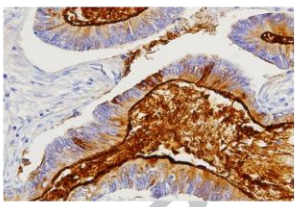
Biliary tract cancers (BTCs) are clinically and genetically heterogeneous. We here analyzed genomic features of 412 BTC samples from Japanese and Italian population by genome sequencing. 32 significantly and commonly mutated genes were identified, some of which negatively affected patient prognosis, including a novel deletion of *MUC17* at *7q22.1*. Cell-of-origin predictions using genetic and epigenetic features suggest hepatocyte-origin of hepatitis-related intrahepatic cholangiocarcinoma. Deleterious germline mutations of cancer-predisposing genes were detected in 11% of BTC patients. BTCs have distinct genetic features including somatic events and germline predisposition.

Genomic features of 412 BTSs

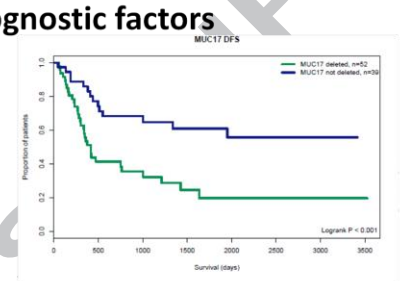
WGS, WES, target seq, CN analysis



TMA

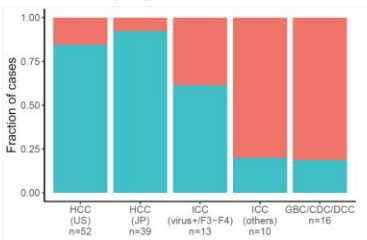


Prognostic factors



Germline variants of cancer predisposing genes

Cell-of-origin prediction by WGS and epigenome



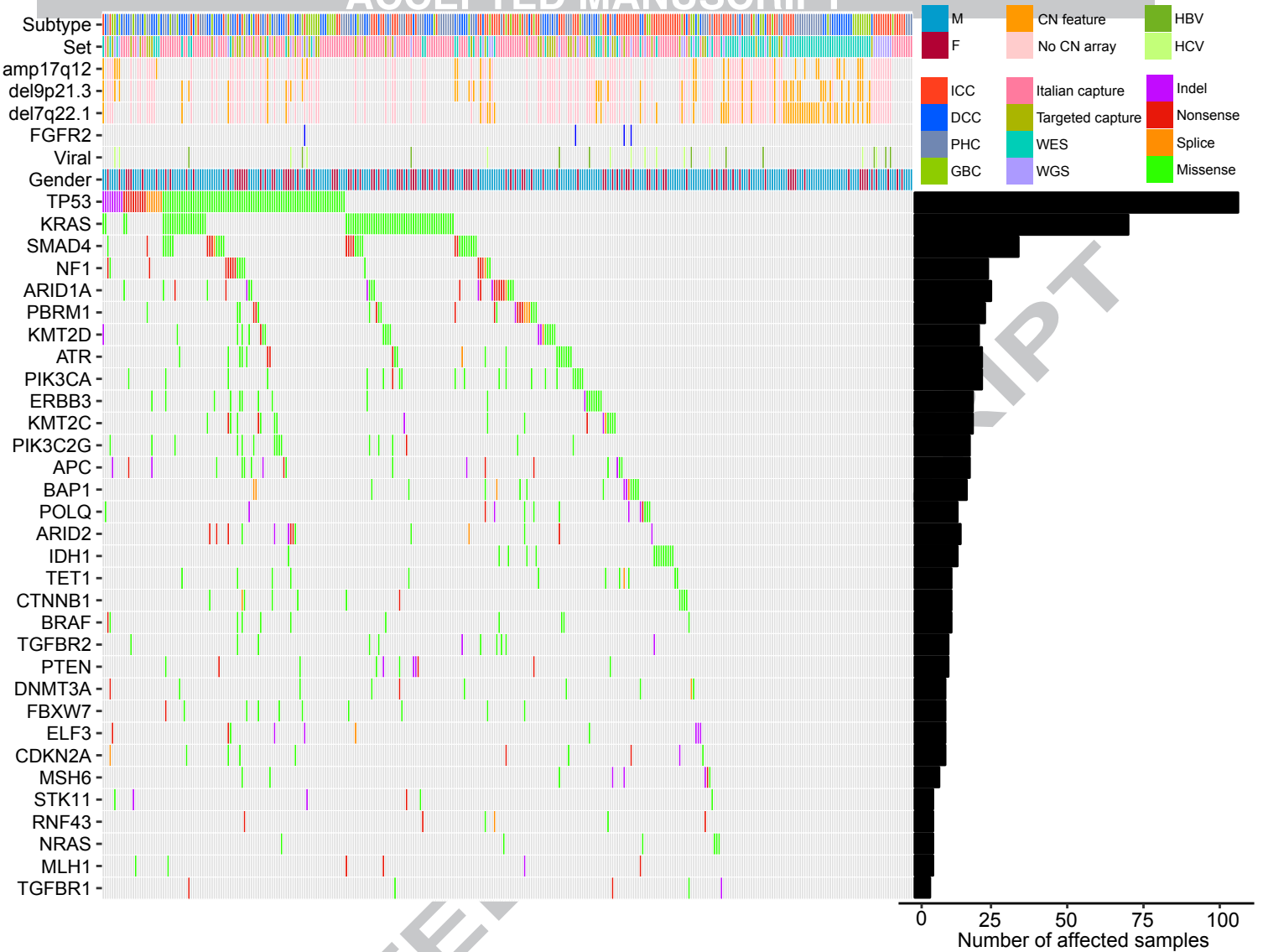
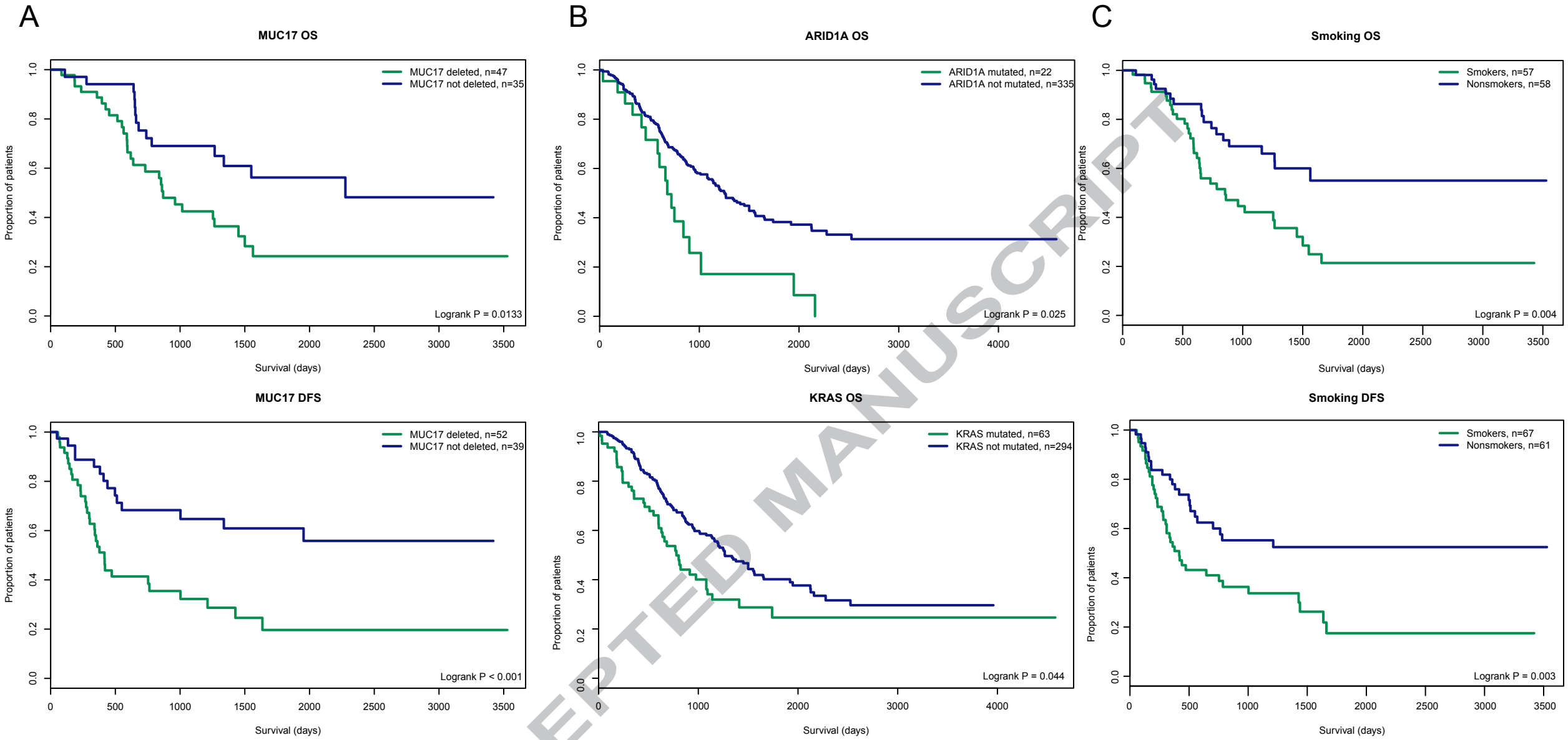


Figure 2



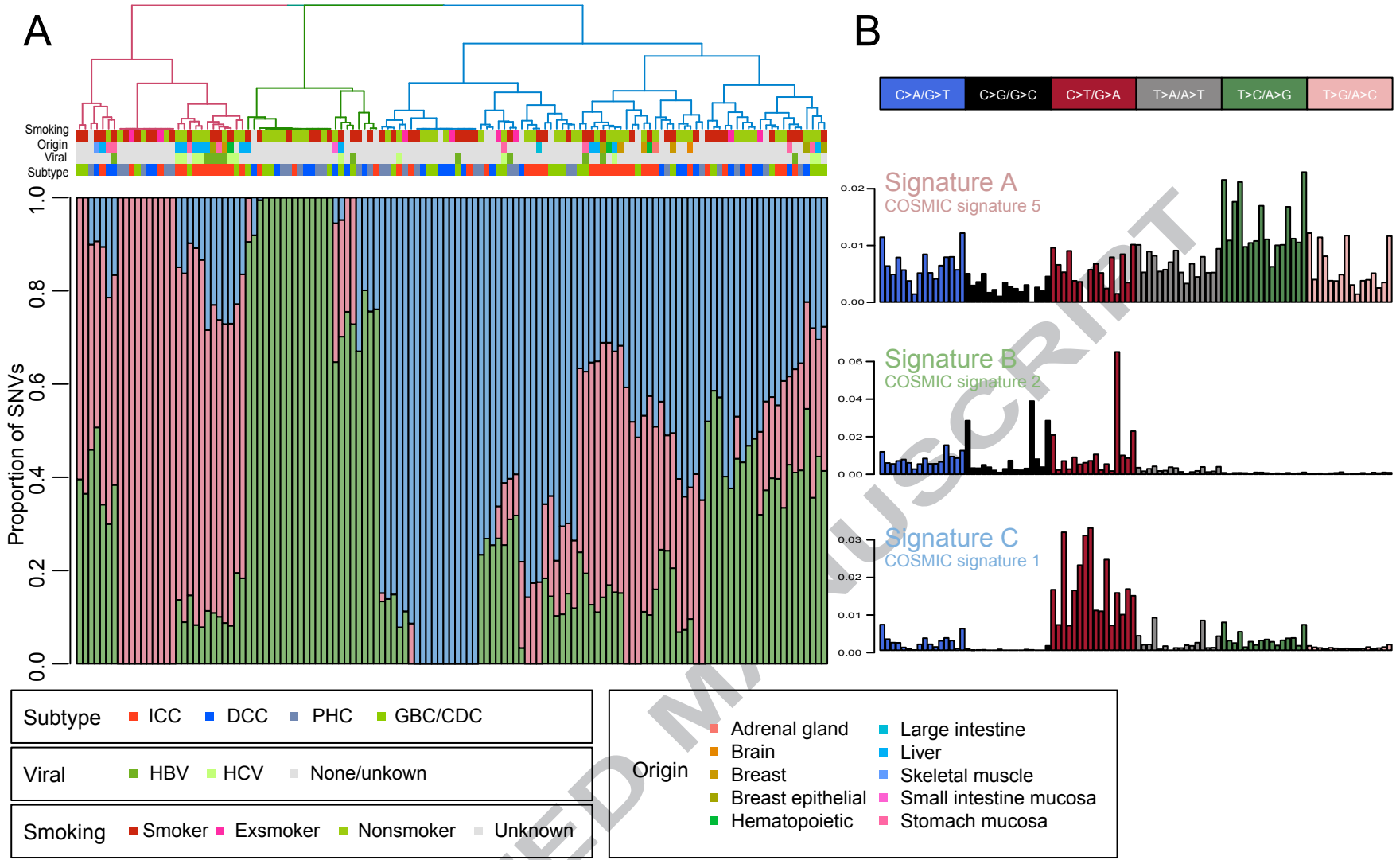


Figure 4