





ORIGINAL ARTICLE

Usefulness of multigene liquid biopsy of bile for identifying driver genes of biliary duct cancers

Shin Ito^{1,2} | Mika Ando³ | Shuichi Aoki³ | Satoshi Soma¹ | Jie Zhang³ |
 Naohiro Hirano³ | Ryosuke Kashiwagi³ | Keigo Murakami⁴ | Shingo Yoshimachi³  |
 Hideaki Sato³ | Akiko Kusaka³ | Masahiro Iseki³ | Koetsu Inoue³ |
 Masamichi Mizuma³ | Kiyoshi Kume⁵ | Kei Nakagawa³  | Atsushi Masamune⁵ |
 Naoki Asano^{6,7} | Jun Yasuda^{1,2,8}  | Michiaki Unno³ 

¹Division of Molecular and Cellular Oncology, Miyagi Cancer Center Research Institute, Natori, Japan

²Division of Cancer Molecular Biology, Tohoku University Graduate School of Medicine, Sendai, Japan

³Department of Surgery, Tohoku University Graduate School of Medicine, Sendai, Japan

⁴Department of Investigative Pathology, Tohoku University Graduate School of Medicine, Sendai, Japan

⁵Division of Gastroenterology, Tohoku University Graduate School of Medicine, Sendai, Japan

⁶Division of Cancer Stem Cell, Miyagi Cancer Center Research Institute, Natori, Japan

⁷Division of Carcinogenesis and Senescence Biology, Tohoku University Graduate School of Medicine, Natori, Japan

⁸Tohoku Medical Megabank Organization, Tohoku University, Sendai, Japan

Correspondence

Shuichi Aoki, Department of Surgery,
 Tohoku University Graduate School of
 Medicine, 1-1 Seiryō-machi, Aoba-ku,
 Sendai 980-8574, Miyagi, Japan.
 Email: shuichi1124@surg.med.tohoku.ac.jp

Jun Yasuda, Division of Molecular and
 Cellular Oncology, Miyagi Cancer Center
 Research Institute, 47-1 Nodayama,
 Aishima-Shiode, Natori, Miyagi 981-1293,
 Japan.
 Email: jun-yasuda@miyagi-pho.jp

Funding information

Japan Society for the Promotion
 of Science, Grant/Award Number:
 21H02996, 21K07111, 21K08748,
 23K08165 and 24K02515

Abstract

Liquid biopsy (LB) is an essential tool for obtaining tumor-derived materials with minimum invasion. Bile has been shown to contain much higher free nucleic acid levels than blood plasma and can be collected through endoscopic procedures. Therefore, bile possesses high potential as a source of tumor derived cell-free DNA (cfDNA) for bile duct cancers. In this study, we show that a multigene panel for plasma LB can also be applied to bile cfDNA for comparing driver gene mutation detection in other sources (plasma and tumor tissues of the corresponding patients). We collected cfDNA samples from the bile of 24 biliary tract cancer cases. These included 17 cholangiocarcinomas, three ampullary carcinoma, two pancreatic cancers, one intraductal papillary mucinous carcinoma, and one insulinoma. Seventeen plasma samples were obtained from the corresponding patients before surgical resection and subjected to the LiquidPlex multigene panel LB system. We applied a machine learning approach to classify possible tumor-derived variants among the prefiltered variant calls by a LiquidPlex analytical package with high fidelity. Among the 17 cholangiocarcinomas, we could detect cancer driver mutations in the bile of

Abbreviations: AC, ampullary carcinoma; BTC, biliary tract cancer; CC, cholangiocarcinoma or gallbladder cancer; cfDNA, cell-free DNA; ddPCR, droplet digital PCR; LB, liquid biopsy; MAF, minor allele frequency; NGS, next-generation sequencing; PC, pancreatic cancer; VAF, variant allele frequency; VCF, variant call format.

Shin Ito and Mika Ando contributed equally to this work.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Author(s). *Cancer Science* published by John Wiley & Sons Australia, Ltd on behalf of Japanese Cancer Association.

10 cases using the LiquidPlex system. Of the biliary tract cancer cases examined with this method, 13 (54%) and 4 (17%) resulted in positive cancer driver mutation detection in the bile and plasma cfDNAs, respectively. These results suggest that bile is a more reliable source for LB than plasma for multigene panel analyses of biliary tract cancers.

KEYWORDS

bile, biliary duct cancer, liquid biopsy, machine learning, multigene panel

1 | INTRODUCTION

Multigene panel LB testing with cfDNA in patient plasma samples has become a routine method in Japanese clinics to identify molecular biomarkers of advanced cancer.¹ This approach is very useful when there are no surgical or biopsy specimens and lack of any indication of biopsy. It also has other potential uses in cancer treatment, including the early detection of recurrence and identification of novel molecular changes that may cause drug resistance. For detecting recurrence, ddPCR analyses or custom-made NGS experiments can be carried out after obtaining the comprehensive somatic mutation profile of a cancer sample. This can be achieved using a set of patient-specific probes designed using the somatic mutation information. It is a well-known fact that the recurrence of cancer is associated with the acquisition of new driver mutations in the cancer cells,^{2,3} hence the multigene panel LB is required to detect new driver mutations.

Biliary tract cancers are malignant neoplasms and associated with a poor prognosis, as they are difficult to diagnose at an early stage.⁴ Because of this, cfDNA in bile is an attractive molecular target for diagnosing BTC. It has been reported that the cfDNA in bile consists of much longer DNA fragments and is found at higher concentrations compared to that in plasma.^{5–7} Endoscopic sampling would be necessary for LB using bile cfDNA, which likely would not be easier than drawing blood. However, bile drainage is an essential part of the initial treatment procedures for BTCs with obstructive jaundice, suggesting that bile sampling would be easy in such cases.^{8,9}

Few reports are available for multigene panel LB for cfDNA in bile (see reviews by Liu et al.⁹ and Arrichiello et al.⁸). In this study, we used a commercially available multigene panel for plasma LB with unique molecular identifiers for detecting cancer driver mutations in bile and plasma cfDNA samples, comparing their mutation detection efficiencies in BTCs. Machine learning for variant calls showed the potential of this approach for detecting pathogenic variants with multigene panel LB.

2 | MATERIALS AND METHODS

2.1 | Samples

From 202X-202X+2, 24 patients with tumors in the pancreatic head lesion invading the bile duct and requiring endoscopic intervention

for biliary drainage were enrolled in the study at Tohoku University hospital. After biliary drainage and bile duct stenting, all patients subsequently underwent surgical resections. As shown in Table 1, the tumors were then pathologically diagnosed as CC (17 cases), AC (three cases), PC (two cases), intraductal papillary mucinous carcinoma (one case), and insulinoma (one case). Among these, exome analyses of tumor tissue samples of 12 cases have been described elsewhere (Ando and Ito et al., unpublished). In addition, we prepared three plasma samples from noncancerous patients to serve as controls for the LiquidPlex analyses (IDT, Coralville, IA, USA), specifically to help determine background noise levels.

2.2 | Bile and plasma cfDNA sample collection

Endoscopic nasobiliary drainage tubes were placed endoscopically for biliary drainage or biopsy of cancerous lesions in the bile duct. We collected 5 mL bile from the endoscopic nasobiliary drainage tubes. Initially, the bile samples were centrifuged at 1600g for 15 min at room temperature, then the supernatants were carefully placed into tubes and stored at -80°C . In 17 of the patients, we collected 5 mL plasma in cfDNA collection tubes (PAXgene Blood ccfDNA Tubes; Qiagen) and stored them at -80°C for later analysis. Cell-free DNA was extracted using a QIAamp Circulating Nucleic Acid Kit (Qiagen) according to the manufacturer's instructions. The amounts of cfDNA are summarized in Table S1.

2.3 | Cell-free DNA library preparation and NGS analysis

Cell-free DNA extracted from patient bile or plasma samples was subjected to LiquidPlex library construction according to the manufacturer's instructions. For this process, 8.2–91.2 ng plasma cfDNA and 79.7–300 ng bile cfDNA were subjected to end-repair, ligation, and addition of molecular barcodes.² The cfDNA libraries were then amplified using two separate PCR steps, with the index sequences incorporated into the DNA fragments in the second PCR. The NGS libraries were quantified using the KAPA Universal Library Quantification Kit (KK4824; Roche) according to the manufacturer's instructions. MiSeq was also used for library quantification on the Illumina platform sequencers.¹⁰ Sequencing data acquisition

TABLE 1 Summary of cases in this study.

Patient ID	Exome	Bile LP	Plasma LP	Classification	Diagnosis	Age (years)	Gender	T	N	M	Stage
#1	YES	BIL3	PL5	CC	Perihilar cholangiocarcinoma	56	Female	2b	1	0	IIIC
#2	YES	BIL15	PL15	CC	Perihilar cholangiocarcinoma	67	Male	2a	2	0	IVA
#3	YES	BIL23	NA	CC	Perihilar cholangiocarcinoma	67	Male	2b	1	0	IIIC
#4	YES	BIL4	PL4	CC	Distal cholangiocarcinoma	82	Male	2	0	0	IIA
#5	YES	BIL5	PL3	CC	Distal cholangiocarcinoma	78	Male	2	0	0	IIA
#6	YES	BIL6	NA	CC	Distal cholangiocarcinoma	73	Female	1	0	0	I
#7	YES	BIL7	PL7	CC	Distal cholangiocarcinoma	90	Male	1	0	0	I
#8	YES	BIL20	PL20	CC	Gallbladder carcinoma	78	Male	1b	0	0	IB
#9	YES	BIL10	PL10	AC	Ampullary carcinoma	80	Male	3b	2	0	IIIB
#10	YES	BIL17	PL17	AC	Ampullary carcinoma	79	Female	3b	2	0	IIIB
#11	YES	BIL21	NA	AC	Ampullary carcinoma	84	Male	1b	0	0	IB
#12	YES	BIL14	PL14	PC	Pancreatic cancer	47	Male	3	1	0	IIB
#14	YES	BIL24	NA	Other	Insulinoma	74	Male	2	0	0	IIA
B101	NA	BIL1	NA	PC	Pancreatic cancer	84	Male	2	0	0	IB
B109	NA	BIL2	PL2	CC	Perihilar cholangiocarcinoma	59	Female	1b	0	0	IB
B110	NA	BIL9	PL9	CC	Intrahepatic cholangiocarcinoma	71	Female	4b	2	1	IVB
B102	NA	BIL11	PL11	CC	Perihilar cholangiocarcinoma	60	Male	4	0	1	IVB
B103	NA	BIL12	PL12	CC	Gallbladder carcinoma	71	Male	2a	0	1	IVB
B104	NA	BIL13	PL13	CC	Perihilar cholangiocarcinoma	70	Male	1b	0	0	IB
B105	NA	BIL16	PL16	CC	Perihilar cholangiocarcinoma	55	Male	2a	0	0	II
B106	NA	BIL18	PL18	CC	Perihilar cholangiocarcinoma	67	Male	4	0	1	IVB
B107	NA	BIL19	NA	CC	Gallbladder carcinoma	82	Female	4	1	0	IVA
B108	NA	BIL22	NA	CC	Perihilar cholangiocarcinoma	77	Female	4	0	0	IIIB
B111	NA	NA	PL8	SPN	Intraductal papillary mucinous carcinoma	39	Female	3	1	0	IIB

Abbreviations: AC, ampullary carcinoma; BIL, bile samples; CC, cholangiocarcinoma or gallbladder cancer; LP, LiquidPlex; NA, not analyzed; PC, pancreatic cancer; PL, plasma; SPN, intraductal papillary mucinous carcinoma.

for mutation analysis was outsourced to Rhelixa, Inc., for which NovaSeq 6000 was used. The 150bp paired-end protocol was used to obtain the cfDNA sequence data.

2.4 | Bile cfDNA ddPCR analysis

The ddPCR Mutation Assay (Bio-Rad) was used to detect the variants called by LiquidPlex. We selected commercially available probes for the somatic mutations identified in the cfDNA samples using ddPCR. Ten probes were selected to verify the mutations detected by LiquidPlex analysis of the bile and plasma cfDNAs. One probe was obtained from Thermo Fisher Scientific: KRAS p.G12V (catalog #A44177, assay ID: Hs000000050_rm). The following probes were obtained from Bio-Rad: NRAS Q61R (dHsaMDS882187944), TP53 R175H (dHsaMDV2010105), TP53 N247I (dHsaMDS684163296), KRAS G12A (dHsaMDV2510586), KRAS A59T (dHsaMDS653784166), NRAS G13R (dHsaMDS295930262), PIK3CA H1047L (dHsaMDS291608817), SMAD4 R361C (dHsaMDS2515076), and BRAF D594G (dHsaMDS280672815) (catalog

#10049047 or 10049550). The ddPCR experiments were carried out using the QX200 AutoDG Droplet Digital PCR IVD system (Bio-Rad). For each ddPCR, 9.9 μ L eluted cfDNA solution (average DNA content = 55.9 ng) was used in quadruplicate for each sample. Bile genomic DNA was extracted as previously described.^{11,12} The total reaction volume was 22 μ L (20,000 drops were generated). The PCR conditions were as follows: initial denaturation at 95°C for 10 min; 40 cycles of 94°C for 30 s and 55°C or 60°C for 1 min; and final denaturation at 98°C for 10 min for enzyme deactivation. The probe fluorophore was FAM/HEX or VIC, and the WT and mutant copy numbers were measured. All ddPCRs were undertaken at least twice. Allele frequencies were calculated using Quanta Software (Bio-Rad) according to the manufacturer's instructions.

2.5 | Bioinformatics of LiquidPlex sequencing: Machine learning process for variant filtering

Sequence read mapping and variant calls were carried out using the Archer Analysis platform (Integrated DNA Technologies, Inc.), a

web-based bioinformatics service for LiquidPlex and other sequencing kits, with default options.² The GRCh37 coordinates served as the reference genome for mapping. The Archer Analysis platform gave us intermediate VCF files, which represent the stepwise filtration of the unreliable variants. Among them, we selected a set of VCF files, *sample_name.combined.filtered.vcfs*, for training and test data of the machine learning process. For the training data, we chose 22 bile or plasma samples that had corresponding cancer exome data (Table 1).

We applied Random Forest¹³ for filtering the variants called by the Archer Analysis platform, as it was previously successfully applied to classify de novo variants of 250 trio (a pair of parents and their children) whole genomic sequencing of a Dutch population.¹⁴ Further details of Random Forest are described in Data S1. Using ANNOVAR,¹⁵ we added annotations for the called variants of a custom-made 54 KJPN variant dataset¹⁶⁻¹⁸ and COSMIC 92,¹⁹ and ClinVar as parameters for machine learning. The 54 KJPN dataset (PMID: 37930845) was downloaded from jMORP (<https://jmorp.megabank.tohoku.ac.jp/downloads>; accessed 4 December 2023), then the bcftools norm-m-option was used to separate multiple allelic sites.^{20,21} Five variants that were confirmed positive by ddPCR were removed from the test data and added to the training data (Figures 1A and S1). After classifying, we annotated the "somatic" variants with ANNOVAR using the AlphaMissense dataset.²²

2.6 | Statistical analysis

Pearson's correlation analysis was used for the cfDNA concentration and LiquidPlex library read count data. Fisher's exact test was used for driver mutation positivity comparisons among the cancer histology samples. The boxplots used for classifying and the allele frequencies were generated with the ggplot2 package in R.

3 | RESULTS

3.1 | Multiplex LB libraries similarly constructed with bile and plasma cfDNAs

As previously described, the bile cfDNA levels were 68.2-fold higher than those of plasma cfDNA in this study (Table S1). We then constructed the LiquidPlex libraries for the bile or plasma cfDNA following the manufacturer's protocols after quantifying the cfDNA using fluorescence (see Materials and Methods) and pooling them with equal volumes in one tube. Using the MiSeq sequencer, we quantified the relative number of DNA fragment copies of each library (see Materials and Methods), but did not observe any significant correlation between the cfDNA concentration and read percentage in a MiSeq run of the pool of libraries

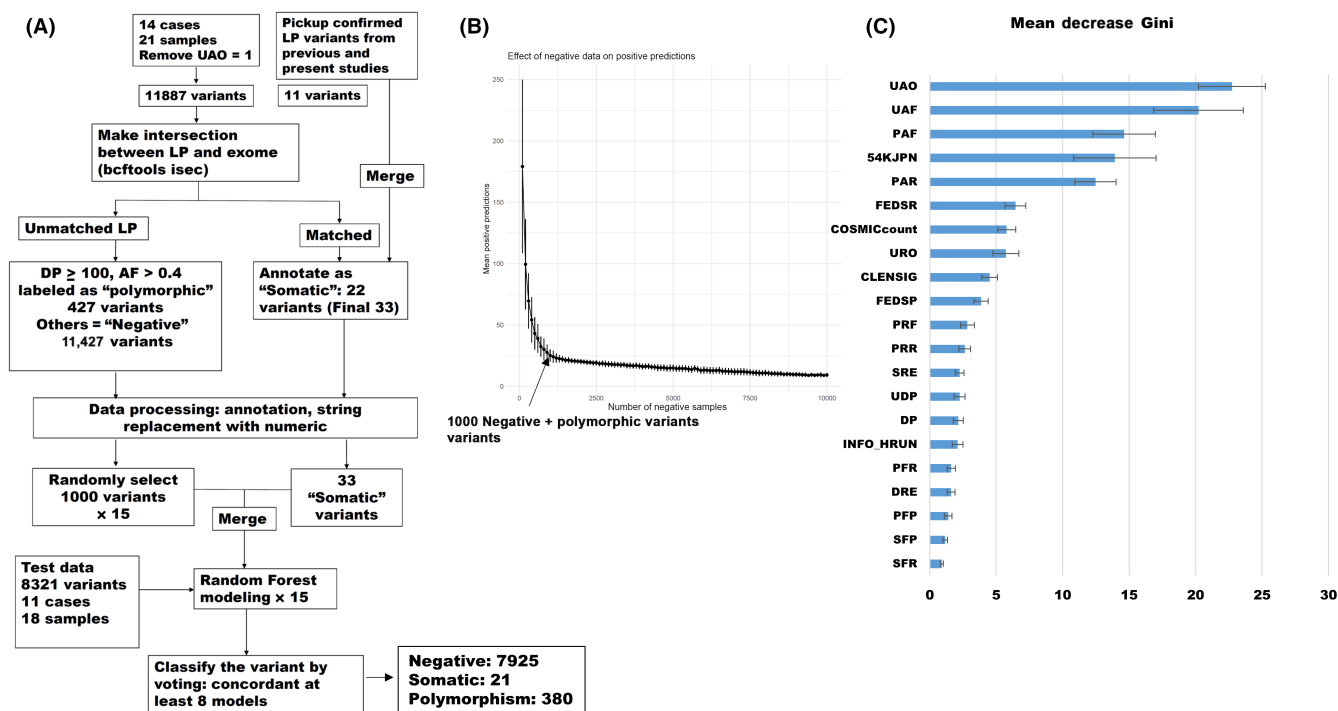


FIGURE 1 Machine learning process to classify the variant calls by multigene panel liquid biopsy (LB). (A) Numbers of variants at each step are indicated near the box. The result of the classification is indicated in the table at the bottom right. (B) Plot of "somatic" variant numbers in Random Forest modeling by changing the "negative and polymorphic" variant numbers for the training data. Vertical and horizontal axes indicate the numbers of variants classified as "somatic" in the test data and numbers of "negative and polymorphic" variants in the training data, respectively. Arrow indicates the number of "negative and polymorphic" variants that we picked up for modeling. (C) Plot of mean decrease in Gini coefficient in this study. The horizontal axis indicates the mean decrease in Gini score and SDs of the scores among the 15 models. The vertical axis indicates the parameters used in the modeling. The parameters were collected by R script.

($r = -0.119$, Table S1). Among the 16 pairs of bile and plasma samples, no correlation was found between bile and plasma for the number of reads in a library ($r = -0.202$, Table S1). The number of library reads moderately correlated with the plasma cfDNA concentration ($r = 0.596$), but not with the bile cfDNA concentration ($r = -0.040$). Neither bile nor plasma cfDNA concentration showed a correlation to the patients' clinical stages ($p = 0.286$ and -0.0688 , respectively; Tables 1 and S1). Considering the experimental principles, it is possible that the plasma cfDNA concentrations were overestimated, especially when the concentration was near the lower limit of measurement (0.2 ng/ μ L). These results suggested that multigene LB libraries were similarly constructed with both bile and plasma cfDNAs, although the plasma cfDNA concentration was much lower than that of bile cfDNA.

3.2 | Machine learning can classify LiquidPlex variant calls by their biological nature

After the LiquidPlex variant call was performed using the Archer Analysis platform, the filtering method of the platform was determined to be too strong. Some of the variants detected by tumor tissue exome analysis of the same patient tumors were missed in the Archer Analysis variant calls, but were detected by ddPCR analysis (Figure S1). Therefore, we applied a machine learning process to filter the variant calls with the Archer Analysis platform. The Random Forest algorithm was used for classifying the LiquidPlex data output. We used the intermediate VCF files (see Materials and Methods), which have 90 parameters described in the header. A summary of the parameters used to build the models and transform the ClinVar significance data into numerical values are summarized in Data S1.

The machine learning workflow is shown in Figure 1A. Our dataset consisted of small numbers of "somatic" variants (33; Table S2), prompting us to require appropriate numbers of "negative and polymorphic" variants for the training dataset. Not doing this would likely result in overfitting to the negative and polymorphic variants, especially if all 11,000 negative variants were used for training. Figure 1B shows the numbers of variants classified as "somatic" in the test dataset by the Random Forest models generated by various numbers of "negative and polymorphic" variants with 33 "somatic" variants in the training dataset. Considering the balance between overfitting to "negative" variants and reduction of noises, 1000 "negative and polymorphic" variants was deemed appropriate (Figure 1B), so these were selected for developing the training datasets.

The average importance of the parameters among the 15 randomly generated models is shown in Figure 1C. Two parameters related to the variant allele frequencies (UAO and UAF, see Data S1) showed higher importance than the other parameters (Figure 1C). The two strand bias-related parameters (PAF and PAR, see Data S1) and 54KJPN were the third to fifth most important parameters in the model, respectively, indicating that the model mainly used parameters related to the variant call qualities (Figure 1C). For public database annotations, 54KJPN showed the highest impact for

classifying in this model. To estimate the precision and recall of our method, we performed predictions of the training data in the same manner and compared the voting results with those of the original classification. The precision and recalls of "somatic" data with individual models were 0.77 and 1.0, respectively. The voting result of the 15 models improved the precision of "somatic" as 0.825, with seven somatic variants mistaken as "polymorphism" or "negative." Among them, five variants were called "polymorphic." Remarkably, the recall of "somatic" classification was 100% and this may reflect that all models used the same positive training dataset. The calling of "negative" showed relatively high precision and recalls, either individual or voting (0.999–1.00). Considering the number of actual negative variation in the training data (more than 11,000), there would still be a high number of errors (14 variants). Notably, most of the false negatives were "polymorphism" (12 of 14 misclassified from negative) and no authentic somatic variants were missed by our model. These results suggested that our method is very reliable for detecting somatic pathogenic variants in cfDNA samples from bile and plasma.

Among the 8316 variants tested, 21 "somatic" variants (0.25%) were detected. In addition, we detected 380 "polymorphic" variants (4.57%) in the tested variants from 18 samples of both bile and plasma cfDNAs (Figure 1A). The comparison of the original Archer Analysis variant call and our Random Forest classification method is summarized in Table S3. The machine learning process improved the plasma cfDNA variant calls significantly: two-fold for the number of driver mutations and three-fold for the mutations matched to the bile cfDNA variants from the same patients. In addition, the variant call quality was improved. The KRAS A59T is a potential variant call artifact because both positives and negatives by Archer Analyses were negative in ddPCR (Figure S2). The variant was eliminated from the variant call (Table S3). Interestingly, we identified a pair of mutations in the TP53 gene for one set of bile and plasma cfDNAs: p.W146R and p.W146 \times fs23 (Figure S3). The Archer Analysis only called the former mutation in the plasma cfDNA, missing the latter base deletion mutation in both the bile and plasma cfDNAs (Table S4).

To evaluate the validity of the model, we compared the labeling of variants and their functional properties (Figure 2). As expected, "somatic" variants were enriched among the pathogenic variants in ClinVar (3.98-fold) and COSMIC entry with more than 10 total cases (3.16-fold) compared with all variants (Figure 2A), suggesting that the "somatic" variants mainly originated from cancer tissues. On the contrary, "polymorphism" variants showed no ClinVar pathogenic variants (Figure 2A). The MAFs in the general population are critical for estimating the biological nature of the corresponding variants.^{22,23} Pathogenic variants usually show lower MAFs in a population, while the presence or absence of polymorphic variants depend on both the MAFs and number of samples. Figure 2A shows the 54KJPN MAF of the three categories classified by the Random Forest model. The polymorphic variants with the higher (>0.4) MAF were enriched 17.2-fold compared to all variants in 54KJPN (Figure 2A). Considering the size of the population (11 cases, 22

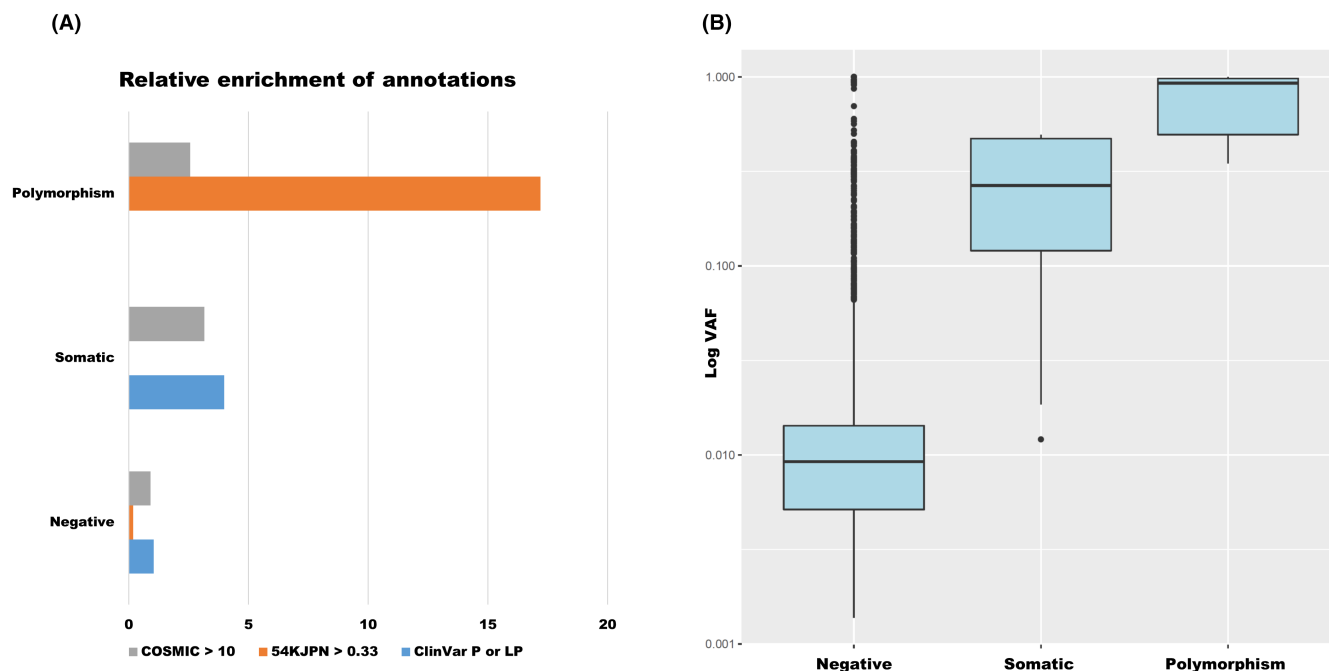


FIGURE 2 Annotation patterns of the variants classified by machine learning. (A) Ratio of variants of three functional annotations for three categories. Vertical axis shows the three categories classified by the Random Forest method (from top to bottom: “Polymorphism,” “Somatic,” and “Negative”). Horizontal axis indicates the relative ratio of the variants in the categories, the ratio of matched variants divided by that of all tested variants. Gray, orange, and blue bars indicate the ratios for COSMIC >10 cases, 54KJPN >0.33, and ClinVar Pathogenic (ClinVar P) and likely_pathogenic (LP), respectively. (B) Categorical box-whisker plots of log of variant allele frequency (VAF) for “negative,” “positive,” and “polymorphism.” X-axis indicates the three categories. Y-axis indicates the log of VAF. Upper and lower ends of the boxes indicate 75% and 25% of each category, respectively. Horizontal black lines in the boxes indicate the averages, while the whiskers indicate the 1.5-fold length of the box vertical plane size of each category. Outliers are shown as dots.

chromosomes/variants), the distribution of MAFs looked reasonable because the probabilities of rare variants (MAF <0.5%) were not high. In terms of “somatic” variants, no such high MAF variants were detected in 54KJPN.

The VAF is the most important parameter for classification (Figure 1C). Figure 2B shows the VAF distribution for three categories. As expected, most of the “polymorphic” variant VAFs were more than 0.4, suggesting that they were of germline origin. However, most of the “somatic” variants were less than 0.3, indicating that these did not originate from the germline. For “negative” variants, the average VAF was nearly 0.001, suggesting that PCR artifacts during data acquisition could be the major cause of the variant calls.

3.3 | Bile LB can detect cancer driver mutations more frequently than plasma LB for BTCs

Figure 3 is an oncoprint for the LB results in this study. We identified 23 different mutations in seven cancer-related genes (Figure 3, Table S4). The most frequently mutated oncogene was the *KRAS* gene (five cases, including one PC). A paralog of *KRAS*, the *NRAS* gene, was also mutated in two cases, with 7 out of 24 cases (29.2%) showing at least one driver mutation of the two *RAS* family genes (Table S4). In addition, two CC cases and one AC case showed possible driver mutations in the *BRAF* gene, indicating that the

RAS-RAF-MAP kinase axis plays a critical role in CCs. Ten out of 24 cases had at least one pathogenic mutation in *ERBB2*, *KRAS*, *NRAS*, and *BRAF*, with these mutations being mutually exclusive (Figure 3). For tumor suppressor genes, nine cases (37.5%) showed a mutation in the *TP53* gene. In addition, we observed pathogenic mutations in the *SMAD4* gene in two cases (Figure 3).

The LiquidPlex LB panel covers only 28 genes and has not been optimized for BTCs. An important function of the multigene panel LB would be detecting circulating and/or cell-free tumor DNA molecules without prior knowledge of the patient's cancer mutations. Table 2 shows that bile LB with the LiquidPlex indicated that 10 out of 17 CC cases were positive for at least one driver mutation, although there was no statistical difference among the CC and other cancerous lesions in the BTCs ($p=0.659$, Fisher's exact test). We also compared the cfDNA detection efficiencies between bile and plasma. Table 2 clearly indicates that bile cfDNA was much more sensitive than plasma cfDNA for detecting BTC driver mutations. Among the 24 cases, 13 (54%) and 4 (17%) cases were positive for driver mutations in LiquidPlex of bile and plasma cfDNAs, respectively.

The low sensitivity of plasma LB for BTCs prompted us to test clear positive controls for LB: germline polymorphisms. Table 3 summarizes the plasma LB variant call results for three high-frequent germline variants among the Japanese population. We could confirm that the presence or absence of the alleles perfectly matched between bile and plasma cfDNAs.

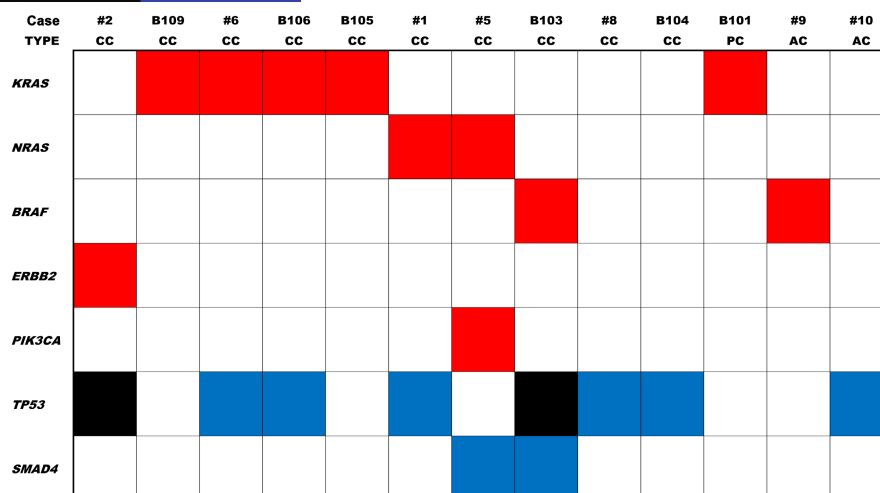


FIGURE 3 Oncoprint of the bile and plasma multigene liquid biopsies (LBs) for biliary tract cancer. Vertical axis indicates the genes in which a driver mutation was detected in this study. Horizontal axis indicates the patients. Results of both the plasma and bile LBs are indicated. Patient identification numbers and tumor types are shown on the top of the oncoprint. Black, blue, and red rectangles indicate the driver mutations as nonsynonymous in tumor suppressor genes, loss of function in tumor suppressor genes, and nonsynonymous in oncogenes, respectively. AC, ampullary carcinoma; CC, cholangiocarcinoma or gallbladder cancer; PC, pancreatic cancer.

Cancer type	No. of cases	Driver-positive cases			
		Bile	%	Plasma	%
Cholangiocarcinoma	17	10	58.8	3	17.6
Ampullary carcinoma	3	2	66.7	1	33.3
Intraductal papillary mucinous carcinoma	1	0	0.0	0	0.0
Insulinoma	1	0	0.0	0	0.0
Pancreatic cancer	2	1	50.00	0	0.0
Total	24	13	54.2	4	16.7

TABLE 2 Detection of cancer driver mutations in bile or plasma liquid biopsies with multigene panel for biliary tract cancers.

3.4 | Machine learning variant classification in bile LB can be confirmed by ddPCR analysis

We next examined whether the variant calls of the multigene LB classified by the machine learning process could be confirmed using ddPCR analysis. Figure 4A shows a representative ddPCR result, specifically *SMAD4* R361C for BIL12. We could confirm the presence of these variants in the bile cfDNAs (Figure 4A). The observed variant fraction in the ddPCR data for bile was lower than that of the final VCF file generated by the Archer Analysis platform (Figure 4B). Interestingly, the plasma cfDNA of the corresponding patient (PL12) showed higher fractions of mutated reads (Figure 4C).

4 | DISCUSSION

In the present study, we found that a multigene panel LB could be applied to bile cfDNAs in BTC cases. A machine learning method for classifying the variants from the LiquidPlex variant calls resulted in 28 possible somatic mutations among more than 8000 variants

detected by the LB analysis. The classification ("somatic" for somatic variants, "negative" for artifacts or uncertain, and "polymorphism" for germline variants) showed expected patterns of annotations with ClinVar, COSMIC, and 54KJPN. Although the panel only consists of 28 genes and was designed for plasma cfDNAs, it could detect cancer driver mutations in nearly two-thirds of the CC cases examined.

Here, we applied a machine learning method to classify the pre-filtered variant calls by the Archer Analysis platform with the default option. We used verified variants with three different types of methods: exome analysis of the tumor tissue genomic DNA, ddPCR analysis of bile and plasma cfDNAs, and clinical testing of the tumor tissue DNA. Most of the "somatic" variants observed in the training data originated from tumor tissue DNA. Recently, the presence of cancer driver variants in tumor and plasma cfDNAs were shown to be highly concordant (see review by Jahangiri and Hurst²⁴). Our results also showed good concordance between the variant classification and clinical annotations (Figure 2), although it is necessary to further verify the classification more extensively.

Using larger panels would potentially improve the observed driver mutation positivity. Even in plasma cfDNA, Mody et al.

showed that cancer driver mutations were positively detected in 76% of BTC cases using the Guardant 360 plasma multigene LB panel, which consists of more than 70 cancer-related genes.²⁵ Several reports have described bile LBs with multigene panels. For example, Shen et al. reported that 8 out of 10 CC cases showed at least one driver mutation in their custom multigene panel with unique molecular identifiers.⁵ Similarly, He et al. undertook a comprehensive study for bile LB with a custom multigene panel using 23 gene mutations and 44 DNA methylation loci, with the results showing 92% sensitivity and 98% specificity.²⁶ Miura et al.

TABLE 3 Genotypes at polymorphic sites in the LiquidPlex target regions.

Variation	TP53:p. P72R	PDGFRA:p. V824V	HRAS:p. H27H
ClinVar significance	drug_ response	Benign	Benign
54KJPN	0.650	0.142	0.170
Bile heterozygous	6	5	1
Bile Alt homozygous	8	1	0
Plasma heterozygous	6	5	1
Plasma Alt homozygous	8	1	0
Cases analyzed	15	15	15

described that 46.5% of BTC cases examined showed positive cancer driver mutation detection using their custom 60 multigene bile LB approach.²⁷ The observed mutation frequencies of *KRAS* and *TP53* in the present study were 20.8% and 37.5%, respectively (Figure 3), which were consistent with data from previous studies.^{6,27–30} Our results suggested that the multigene LB can be readily applied. Additionally, bile cfDNA was a significantly better source for detecting cancer driver mutations in BTCs than plasma cfDNA. However, plasma cfDNA did show higher signal for mutations than bile cfDNA in one case (Figure 4B,C). Further studies are needed to establish the clinical significance of the bile LB for BTC treatment development.

In addition to the small numbers of the target genes, there are several limitations in the present study, especially the relatively small number of samples. Actually, we failed to identify pathogenic variants that frequently are observed in the BTCs, such as *IDH1* p.R132C. It would be possible that the machine learning method we applied is sensitive to the “observed” correct variants in the training data and might not be sensitive enough for the variants that were not in the training data. It is essential to use larger and more heterogeneous training data for the LiquidPlex multigene panel for clinical use.

The concordance among variant calls between bile and plasma cfDNA LiquidPlex data for our 16 cases (Tables 1 and S4) was not

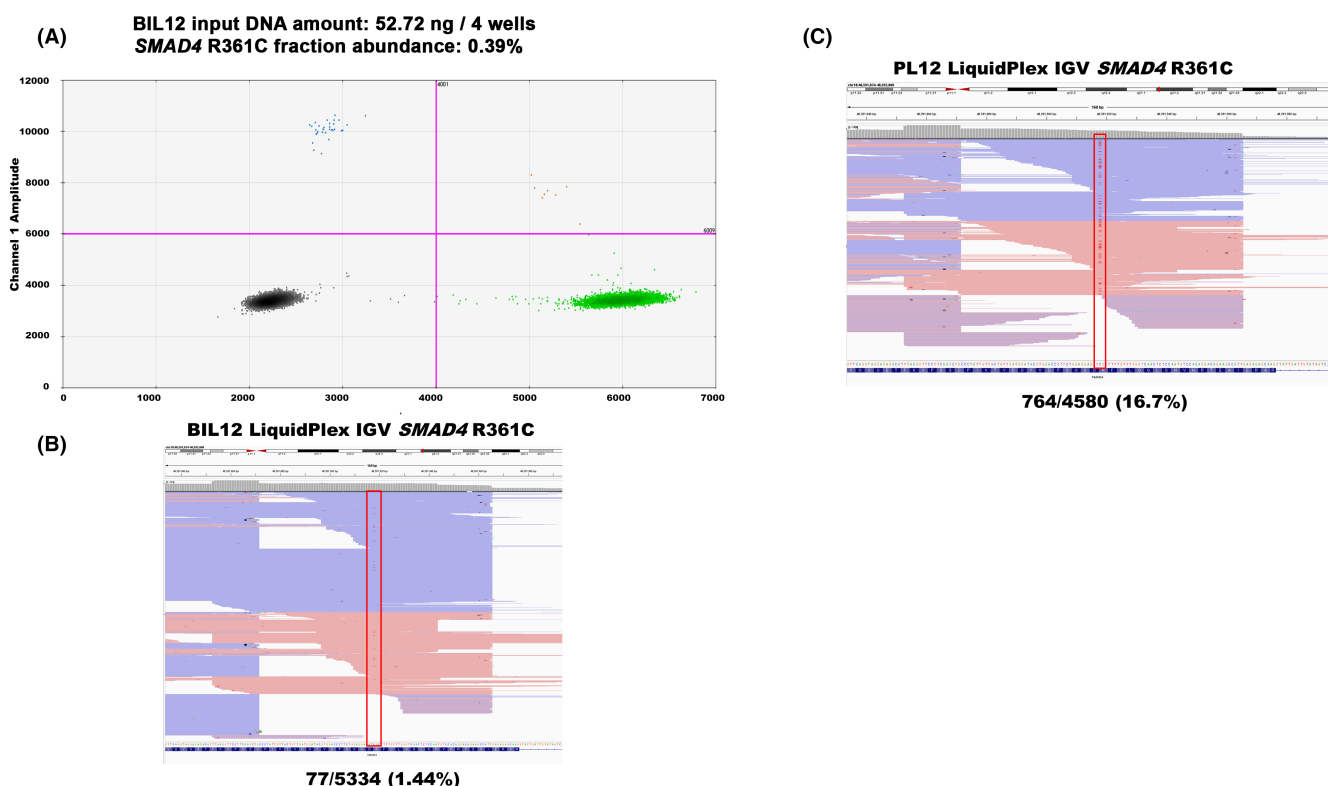


FIGURE 4 Droplet digital PCR (ddPCR) analysis of bile cell-free DNA (cfDNA) samples. (A) ddPCR scatter plot for *SMAD4* R361C cfDNAs in BIL12. Vertical axes indicate the VIC signals (mutant), while the horizontal axes indicate the FAM/HEX signals (WT). Magenta lines in the planes indicate the positive signal thresholds. Dots indicate the FAM/HEX and VIC signals of pixels with the colors of the genotype decisions (blue, mutated; green, WT; orange, heterozygous; black, no signal). (B) Corresponding Integrative Genomics Viewer (IGV) images of the BIL12 cfDNA LiquidPlex data. (C) Corresponding IGV images of the PL12 cfDNA LiquidPlex data.

high (30%, Table S4) compared with previous reports, such as those by Gou et al. and Driescher et al.^{31,32} Considering the observed concordance between the bile LB and exome analysis data, the main cause of the discrepancies between the bile and plasma LB variant calls is possibly the lower amount of plasma cfDNA in our samples for variants positive only in bile. For plasma-specific variants in LB, Driescher et al. suggested the possibility of clonal hematopoiesis.³² Further studies are needed to confirm to what extent clonal hematopoiesis influences our bile LB results.

AUTHOR CONTRIBUTIONS

Shin Ito: Data curation; investigation; methodology; validation; visualization. **Mika Ando:** Data curation; investigation; methodology. **Shuichi Aoki:** Conceptualization; formal analysis; methodology; project administration; visualization; writing – original draft. **Satoshi Soma:** Validation. **Jie Zhang:** Resources. **Naohiro Hirano:** Resources. **Ryosuke Kashiwagi:** Resources. **Keigo Murakami:** Resources; writing – review and editing. **Shingo Yoshimachi:** Resources. **Hideaki Sato:** Resources. **Akiko Kusaka:** Resources. **Masahiro Iseki:** Resources. **Koetsu Inoue:** Resources. **Masamichi Mizuma:** Resources; writing – review and editing. **Kei Nakagawa:** Funding acquisition; resources. **Kiyoshi Kume:** Resources. **Atsushi Masamune:** Investigation; methodology. **Naoki Asano:** Supervision; writing – review and editing. **Jun Yasuda:** Conceptualization; formal analysis; funding acquisition; methodology; project administration; visualization; writing – original draft. **Michiaki Unno:** Funding acquisition; project administration; writing – review and editing.

ACKNOWLEDGMENTS

The authors express their deep gratitude to all surgeons and physicians who performed the endoscopic interventions and resections at Tohoku University Hospital. We thank J. Iacona, Ph.D., from Edanz for editing a draft of this manuscript.

FUNDING INFORMATION

This study was supported by JSPS KAKENHI (grant numbers: 21K07111 [to J. Yasuda], 24K02515 and 21H02996 [to M. Unno], 21K08748 [to K. Nakagawa], and 23K08165 [to S. Aoki]).

CONFLICT OF INTEREST STATEMENT

Michiaki Unno is an editorial board member of *Cancer Science*. The other authors declare no conflict of interest.

ETHICS STATEMENT

Approval of the research protocol by an institutional review board: The study was approved by the Ethics Committee of Tohoku University Graduate School of Medicine (#2023-1-487). This study was conducted in accordance with the principles of the Declaration of Helsinki.

Informed consent: Written informed consent was obtained from all participants.

Registry and the registration no. of the study/trial: N/A.

Animal studies: N/A.

ORCID

Shingo Yoshimachi  <https://orcid.org/0000-0002-8823-6696>

Kei Nakagawa  <https://orcid.org/0000-0002-3058-5674>

Jun Yasuda  <https://orcid.org/0000-0002-3887-6871>

Michiaki Unno  <https://orcid.org/0000-0002-2145-6416>

REFERENCES

- Horie S, Saito Y, Kogure Y, et al. Pan-cancer comparative and integrative analyses of driver alterations using Japanese and international genomic databases. *Cancer Discov*. 2024;14:786-803.
- Ito S, Tsurumi K, Shindo N, et al. Multi-gene liquid biopsy to detect resistance to first-line osimertinib in patients with EGFR-mutated lung adenocarcinoma. *Anticancer Res*. 2023;43:5031-5040.
- Shaw AT, Friboulet L, Leshchiner I, et al. Resensitization to Crizotinib by the Lorlatinib ALK resistance mutation L1198F. *N Engl J Med*. 2016;374:54-61.
- Miyazaki M, Yoshitomi H, Miyakawa S, et al. Clinical practice guidelines for the management of biliary tract cancers 2015: the 2nd English edition. *J Hepatobiliary Pancreat Sci*. 2015;22:249-273.
- Shen N, Zhang D, Yin L, et al. Bile cell-free DNA as a novel and powerful liquid biopsy for detecting somatic variants in biliary tract cancer. *Oncol Rep*. 2019;42:549-560.
- Han JY, Ahn KS, Kim TS, et al. Liquid biopsy from bile-circulating tumor DNA in patients with biliary tract cancer. *Cancers (Basel)*. 2021;13:4581.
- Shen N, Zhu B, Zhang W, et al. Comprehensive evaluation and application of a novel method to isolate cell-free DNA derived from bile of biliary tract cancer patients. *Front Oncol*. 2022;12:891917.
- Arrichiello G, Nacca V, Paragliola F, Giunta EF. Liquid biopsy in biliary tract cancer from blood and bile samples: current knowledge and future perspectives. *Explor Target Antitumor Ther*. 2022;3:362-374.
- Liu F, Hao X, Liu B, Liu S, Yuan Y. Bile liquid biopsy in biliary tract cancer. *Clin Chim Acta*. 2023;551:117593.
- Katsuoka F, Yokozawa J, Tsuda K, et al. An efficient quantitation method of next-generation sequencing libraries by using MiSeq sequencer. *Anal Biochem*. 2014;466:27-29.
- Ito S, Sato I, Mochizuki M, et al. Robustness of a cancer profiling test using formalin-fixed paraffin embedded tumor specimens. *Anticancer Res*. 2021;41:1341-1348.
- Miyabe S, Ito S, Sato I, et al. Clinical and genomic features of non-small cell lung cancer occurring in families. *Thorac Cancer*. 2023;14:940-952.
- Breiman L. Random Forest. *Mach Learn*. 2001;45:5-32.
- Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet*. 2014;46:818-825.
- Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38:e164.
- Tadaka S, Saigusa D, Motoike IN, et al. jMorp: Japanese Multi Omics Reference Panel. *Nucleic Acids Res*. 2018;46:D551-D557.
- Tadaka S, Katsuoka F, Ueki M, et al. 3.5KJPNv2: an allele frequency panel of 3552 Japanese individuals including the X chromosome. *Hum Genome Var*. 2019;6:28.
- Tadaka S, Kawashima J, Hishinuma E, et al. jMorp: Japanese Multi-Omics Reference Panel update report 2023. *Nucleic Acids Res*. 2024;52:D622-D632.
- Tate JG, Bamford S, Jubb HC, et al. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res*. 2019;47:D941-D947.
- Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011;27:2987-2993.

21. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *Gigascience*. 2021;10:giab008.
22. Cheng J, Novati G, Pan J, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*. 2023;381:eadg7492.
23. Tokunaga H, Iida K, Hozawa A, et al. Novel candidates of pathogenic variants of the *BRCA1* and *BRCA2* genes from a dataset of 3,552 Japanese whole genomes (3.5KJPNv2). *PLoS One*. 2021;16:e0236907.
24. Jahangiri L, Hurst T. Assessing the concordance of genomic alterations between circulating-free DNA and tumour tissue in cancer patients. *Cancers (Basel)*. 2019;11:1938.
25. Mody K, Kasi PM, Yang J, et al. Circulating tumor DNA profiling of advanced biliary tract cancers. *JCO Precis Oncol*. 2019;3:1-9.
26. He KY, Li X, Kelly TN, et al. Leveraging linkage evidence to identify low-frequency and rare variants on 16p13 associated with blood pressure using TOPMed whole genome sequencing data. *Hum Genet*. 2019;138:199-210.
27. Miura Y, Ohyama H, Mikata R, et al. The efficacy of bile liquid biopsy in the diagnosis and treatment of biliary tract cancer. *J Hepatobiliary Pancreat Sci*. 2024;31:329-338.
28. He S, Zeng F, Yin H, et al. Molecular diagnosis of pancreaticobiliary tract cancer by detecting mutations and methylation changes in bile samples. *EClinicalMedicine*. 2023;55:101736.
29. Guo Q, Lakatos E, Bakir IA, Curtius K, Graham TA, Mustonen V. The mutational signatures of formalin fixation on the human genome. *Nat Commun*. 2022;13:4487.
30. Kinugasa H, Nouse K, Ako S, et al. Liquid biopsy of bile for the molecular diagnosis of gallbladder cancer. *Cancer Biol Ther*. 2018;19:934-938.
31. Gou Q, Zhang CZ, Sun ZH, et al. Cell-free DNA from bile outperformed plasma as a potential alternative to tissue biopsy in biliary tract cancer. *ESMO Open*. 2021;6:100275.
32. Driescher C, Fuchs K, Haeberle L, et al. Bile-based cell-free DNA analysis is a reliable diagnostic tool in Pancreatobiliary cancer. *Cancers (Basel)*. 2020;13:39.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Ito S, Ando M, Aoki S, et al.

Usefulness of multigene liquid biopsy of bile for identifying driver genes of biliary duct cancers. *Cancer Sci*.

2024;115:4054-4063. doi:[10.1111/cas.16365](https://doi.org/10.1111/cas.16365)