






## Ultra-Deep Sequencing of Plasma-Circulating DNA for the Detection of Tumor- Derived Mutations in Patients with Nonmetastatic Colorectal Cancer

Huu-Thinh Nguyen, Bac An Luong, Duc-Huy Tran, Trong-Hieu Nguyen, Quoc Dat Ngo, Linh Gia Hoang Le, Quoc Chuong Ho, Hue-Hanh Thi Nguyen, Cao Minh Nguyen, Vu Uyen Tran, Truong Vinh Ngoc Pham, Minh Triet Le, Ngoc An Trinh Le, Trung Kien Le, Thanh Luan Nguyen, Hong-Anh Thi Pham, Hong Thuy Le, Hong Diep Thi Duong, Anh Vu Hoang, Hoang Bac Nguyen, Kiet Truong Dinh, Minh-Duy Phan, Hoai-Nghia Nguyen, Thanh-Thuy Thi Do, Hoa Giang, Le Son Tran & Diep Tuan Tran


To cite this article: Huu-Thinh Nguyen, Bac An Luong, Duc-Huy Tran, Trong-Hieu Nguyen, Quoc Dat Ngo, Linh Gia Hoang Le, Quoc Chuong Ho, Hue-Hanh Thi Nguyen, Cao Minh Nguyen, Vu Uyen Tran, Truong Vinh Ngoc Pham, Minh Triet Le, Ngoc An Trinh Le, Trung Kien Le, Thanh Luan Nguyen, Hong-Anh Thi Pham, Hong Thuy Le, Hong Diep Thi Duong, Anh Vu Hoang, Hoang Bac Nguyen, Kiet Truong Dinh, Minh-Duy Phan, Hoai-Nghia Nguyen, Thanh-Thuy Thi Do, Hoa Giang, Le Son Tran & Diep Tuan Tran (2021): Ultra-Deep Sequencing of Plasma-Circulating DNA for the Detection of Tumor- Derived Mutations in Patients with Nonmetastatic Colorectal Cancer, *Cancer Investigation*, DOI: [10.1080/07357907.2021.2017951](https://doi.org/10.1080/07357907.2021.2017951)

To link to this article: <https://doi.org/10.1080/07357907.2021.2017951>

 View supplementary material 

 Published online: 22 Dec 2021.

 Submit your article to this journal 



 Article views: 28

 View related articles 

 View Crossmark data 



## Ultra-Deep Sequencing of Plasma-Circulating DNA for the Detection of Tumor-Derived Mutations in Patients with Nonmetastatic Colorectal Cancer

Huu-Thinh Nguyen<sup>a</sup>, Bac An Luong<sup>b</sup>, Duc-Huy Tran<sup>a</sup>, Trong-Hieu Nguyen<sup>c</sup>, Quoc Dat Ngo<sup>b</sup>, Linh Gia Hoang Le<sup>b</sup>, Quoc Chuong Ho<sup>b</sup>, Hue-Hanh Thi Nguyen<sup>c</sup>, Cao Minh Nguyen<sup>c</sup>, Vu Uyen Tran<sup>c</sup>, Truong Vinh Ngoc Pham<sup>a</sup>, Minh Triet Le<sup>a</sup>, Ngoc An Trinh Le<sup>a</sup>, Trung Kien Le<sup>a</sup>, Thanh Luan Nguyen<sup>c</sup>, Hong-Anh Thi Pham<sup>c</sup>, Hong Thuy Le<sup>b</sup>, Hong Diep Thi Duong<sup>b</sup>, Anh Vu Hoang<sup>b</sup>, Hoang Bac Nguyen<sup>a</sup>, Kiet Truong Dinh<sup>c</sup>, Minh-Duy Phan<sup>c</sup> , Hoai-Nghia Nguyen<sup>b</sup>, Thanh-Thuy Thi Do<sup>c</sup>, Hoa Giang<sup>c</sup>, Le Son Tran<sup>c</sup> , and Diep Tuan Tran<sup>b</sup>

<sup>a</sup>University Medical Center, Ho Chi Minh City, Vietnam; <sup>b</sup>University of Medicine and Pharmacy, Ho Chi Minh City, Vietnam; <sup>c</sup>Medical Genetics Institute, Ho Chi Minh City, Vietnam

### ABSTRACT

Identification of tumor-derived mutation (TDM) in liquid biopsies (LB), especially in early-stage patients, faces several challenges, including low variant-allele frequencies, interference by white blood cell (WBC)-derived mutations (WDM), benign somatic mutations and tumor heterogeneity. Here, we addressed the above-mentioned challenges in a cohort of 50 non-metastatic colorectal cancer patients, via a workflow involving parallel sequencing of paired WBC- and tumor-gDNA. After excluding potential false positive mutations, we detected at least one TDM in LB of 56% (28/50) of patients, with the majority showing low-patient coverage, except for one TDM mapped to *KMT2D* that recurred in 30% (15/30) of patients.

### ARTICLE HISTORY

Received 16 July 2021  
Revised 20 October 2021  
Accepted 9 December 2021

### KEYWORDS

Cell-free DNA; circulating tumor DNA; early detection; colorectal cancer; tumor-specific mutations




### Introduction


Colorectal cancer (CRC) remains the third-most common cancer, with 1.85 million persons afflicted, worldwide, and the second highest cause of cancer mortality, at 881,000 deaths per year (1). This is despite screening methods, such as colonoscopy and sigmoidoscopy, for those at higher risk, including family history and age over 50 (2). Those approaches have a miss rate of 5% and remain excessively expensive for low- and middle-income nations, in addition to being invasive, with possible complications such as perforation (2,3). Also, while localized disease has a 91% five-year survival rate, that number falls to 14% for distant metastases (4), thus underscoring the need for biomarkers for nonmetastatic CRC.

To improve ease of cancer detection and monitoring, it has been found that capture of circulating tumor DNA (ctDNA, i.e., “liquid biopsies”) is feasible, particularly for metastatic disease, which

possesses myriad mutations (5–8). These are relatively noninvasive, can capture tumor heterogeneity, and readily detect residual disease (9,10). Indeed, one model predicts that the rate of tumor shedding of DNA (presumably via apoptosis or necrosis) would allow detection of 40% lower-sized lesions, and relapsed disease up to 140 days earlier, compared to current imaging methodologies (10). Moreover, the U.S. Food and Drug Administration has now approved two such liquid biopsies for metastatic cancers, including a 55-mutant gene set for non-small cell lung cancer and a 324-mutant gene assay for prostate cancer, using massive parallel sequencing (MPS) to detect mutations (11).

Mutation-based detection assays of ctDNA have been developed by several groups to detect cancers in asymptomatic individuals, while they are still curable. These studies showed that sensitivity increases with increasing stage, i.e., from

**CONTACT** Le Son Tran  [leson1808@gmail.com](mailto:leson1808@gmail.com)  University Medical Center, Ho Chi Minh City, Vietnam; Diep Tuan Tran [dieptuan@ump.edu.vn](mailto:dieptuan@ump.edu.vn)   
University of Medicine and Pharmacy, Ho Chi Minh City, Vietnam

 Supplemental data for this article can be accessed [here](#).

approximately 40% in stage I to approximately 80% in stage III malignancy (5,7,12,13). It is thought that the fraction ctDNA, and the number of alterations in cancer cells, are the key factors in the detection of ctDNA. In cancer patients, ctDNA generally represents a small proportion of all cfDNA, ranging from  $\geq 5$ –10%, in late-stage disease, to  $\leq 0.01$ –1.0%, in early-stage disease (6). This leads to low variant-allele frequencies (VAFs) of tumor-derived mutations, referring to the percentage of sequence reads matching specific DNA variants, divided by the overall coverage at those loci (14). Therefore, a highly sensitive technique is required to capture such mutations in cfDNA. For instance, using a targeted sequencing method (“TECseq”), Phallen et al. (12) detected somatic mutations in the plasma of 71% colorectal, 59% breast, 59% lung, and 68% ovarian cancers from 200 patients with stage I or II disease. Alternatively, to maximize the sensitivity and specificity of early cancer detection, Cohen et al. (5) developed a multi-analyte blood test (“CancerSeek”) by simultaneously detecting mutations in 16 cancer-related genes, combined with circulating protein biomarkers. The assay achieved median sensitivities of 43%, 73%, and 78% for stage I, II and III disease, respectively, in screening for 8 common cancer types (including CRC). Recently, the biological properties of ctDNA such as the shortening fragment sizes, have been exploited to improve the accuracy of mutation-based detection of ctDNA in early-stage cancer and low disease burden. Cristiano et al. (15) showed that the test based on the combination of DNA fragmentation patterns and mutations could enhance the sensitivity for cancer early detection. Another challenge is the fact that some variants could be due to clonal hematopoiesis of blood cells, which must be filtered to identify circulating, tumor-unique mutations (16). Moreover, somatic mutations from healthy subjects, which are clearly unrelated to the presence of malignant disease, can interfere with the identification of cancer-specific mutations (17–19). As we address in this report, however, none of these early cancer detection tests are clinically approved for cancer screening.

To address these challenges, for CRC, we procured a valuable and unique set of patient

samples, consisting of 50 cases (stage 0 to stage IIIA disease) and 96 healthy controls. Moreover, for greater sensitivity and accuracy, we used ultra-deep MPS, including the addition of unique molecular identifiers (UMIs), allowing suppression of sequencing error and detection of variants at VAFs  $\leq 0.001$  (20). In our cohort, removal of hematopoietic and healthy control variants revealed a set of 33 tumor-derived mutations, with identification of 10 common CRC mutant genes, with the tumor suppressor *APC* being the most frequent. We hold that this study provides proof-of-principle for eventual clinical employment of circulating DNA, via liquid biopsy, for detection of nonmetastatic colorectal cancer.

## Materials and methods

### Patient recruitment

A total of 50 patients with colorectal cancer (CRC) and 96 healthy subjects from the University of Medicine and Pharmacy at Ho Chi Minh City, Vietnam were recruited to this study. The recruitment criteria for CRC patients were early stage (stage I, II) or showed nonmetastatic disease (stage IIIA) and naivety to treatment. Healthy individuals were those who showed no signs of colorectal diseases at the time of colonoscopy. Comprehensive details of patients’ clinical factors are summarized in Table S1. This study was approved by the Ethics Committee of the University of Medicine and Pharmacy at Ho Chi Minh City, Vietnam. All patients were given written informed consent prior to participation in the study.

### Clinical sample collection and DNA isolation

Paired formalin-fixed paraffin embedded (FFPE) tumor tissues and blood were collected from 50 CRC patients. Prior to tissue biopsy, 10 mL of peripheral blood was drawn in  $K_2$ -EDTA tubes (BD Vacutainer, USA), stored at room temperature for a maximum of 4 hours, followed by 2 rounds of centrifugation ( $2,000 \times g$  for 10 min and then  $16,000 \times g$  for 10 min) to separate plasma from buffy coat fractions (white blood cells). The plasma (4–6 mL) and buffy coats were then collected and stored at  $-80^\circ\text{C}$  until DNA

extraction. Cell-free DNA (cfDNA) from 2 ml of plasma, and genomic DNA (gDNA) from buffy coats, were isolated using a MagMAX Cell-Free DNA Isolation kit (Thermo Fisher, USA), following the manufacturer's instructions.

Tumor-rich areas, defined as those containing at least 50% of tumor cells identified by hematoxylin and eosin (H&E) staining, of the FFPE tissues, were micro-dissected. Tumor tissue-derived DNA was then extracted using QIAamp DNA FFPE Tissue Kit (Qiagen, Hilden, Germany), following the manufacturer's instructions.

Isolated gDNA from both FFPE samples and white blood cells (WBCs) were fragmented using a NEBNext dsDNA Fragmentase (New England Biolabs, USA), following the manufacturers' instructions. Sheared gDNA was subject to size selection for 100–1000 bp fragments using KAPA Pure Beads (Roche, Switzerland). Both cell-free DNA from plasma, and gDNA from FFPE or WBCs, were then quantified using the QuantiFluor dsDNA system (Promega, USA) and Quantus Fluorometer (Promega).

#### **Ultra-deep massively parallel sequencing with unique molecular identifier tagging**

A minimum of 1 ng cfDNA was subject to sequencing library preparation with unique molecular identifier (UMI) tagging (21), using the ThruPLEX Tag-seq Kit (Takara Bio, USA), following the manufacturer's instructions. Equal amounts of libraries (100 ng per sample) were then pooled together and hybridized to an xGen Lockdown probe panel targeting the 20 most frequently mutated genes in CRC, as selected from the public database COSMIC (Catalogue of Somatic Mutations in Cancer) (Table S2) (22).

From FFPE and WBC specimens, libraries were prepared from 30 ng gDNA using the NEBNext Ultra II FS DNA library prep kit (New England Biolabs, USA), following the manufacturer's instructions. Like cfDNA, libraries were pooled before hybridization with the xGen Lockdown probes. After hybridization, sequencing was run using MGI DNBSEQ Sequencing Technology (BGI, China), at deep coverage of 20 million PE100 reads per cfDNA- or FFPE- gDNA

sample. For gDNA libraries from WBCs, sequencing was carried out at 5 million PE (paired end) 100 reads per sample.

#### **Variant calling using Vardict**

Each sample was barcoded with a single 8-bp index in the P7 primer, and each DNA fragment was tagged with a UMI consisting of a random 6-bp sequence, at both ends. FastQC Version 0.11.9 (Illumina) was used to check the quality of the sequenced reads, followed by trimming using Trimmomatic (23), to yield 75-bp reads. Pair-end (PE) reads, and their corresponding UMI sequences, were generated using the bcl2fastq package version 2.20.0.422 (Illumina). The reads were aligned to the human genome (hg38), using the Burrows-Wheeler Aligner (BWA) package (24), and then grouped by their UMIs to determine a consensus sequence for each fragment, eliminating sequencing and PCR errors. The Fgbio version 1.3.0 package was used to call UMI consensus reads, which were then grouped by their UMI tag (25). Reads with UMIs shorter than 8-bp were excluded. Consensus reads were called on the group with more than 3 identical reads, using Fgbio CallMolecularConsensusReads (26), and BWA used to align consensus reads to the human genome (hg38) (24). AstraZeneca's Vardict version 1.8 was used to call each variant (27), and annotations attached using the Ensembl Variant Effect Predictor (28).

To minimize the false-positive results, we used high thresholds for calling mutations as tumor-derived mutations (TDMs). Specifically, a mutation identified in cfDNA was considered as a tumor-derived mutation (TDM) only when: (1) it did not overlap with mutations detected in paired WBC samples; (2) it was concordantly detected in paired tissues samples; and (3) it did not overlap with mutations in healthy controls or had significantly higher frequency in the cancer cohort than that in the control cohort.

#### **Statistical analysis**

The Mann-Whitney test was used to compare the median age of CRC patients and healthy subjects. Chi-squared ( $\chi^2$ ) test was performed to compare

gender and mutation frequencies between the CRC and control cohorts, and the  $p$ -values were subsequently adjusted by Bonferroni correction. Correlations between mutational loads and tumor size were made using Spearman's rank test, while Pearson's correlation coefficient test was used to assess correlations between WBC-derived mutational VAFs in liquid biopsies and WBC samples. All statistical analyses were carried out using Python (v3.7) with some common data analysis packages: numpy, scipy, pandas. All visualizations are created with the help of matplotlib, pyplot and pyoncoprint.

## Results

### Patients' clinical features

Paired samples of tissue (FFPE) biopsies and blood samples were collected from 50 patients diagnosed with CRC by clinical histology, all of whom were confirmed to be naive to treatment at the time of sample collection. Parallely, blood samples were collected from 96 healthy individuals (i.e., no colon lesions found by colonoscopy) (Figure S1A). Patients in the CRC cohort had a higher median age compared to controls (63 versus 34,  $p < 0.001$ ), and a higher ratio of females to males (36% female –64% male vs 63.7% female– 30.8% male, Table 1). Of 50 CRC patients, 7 (14%) patients were stage 0-I, and 20 (40%) and 21 (42%) patients were stage II and stage IIIA, respectively. Two (4%) patients were diagnosed with non-metastatic CRC, but lacked information on clinical stage (Table 1). Of the 50, 35 patients (70%) had tumors in the colon, and the remaining 15 patients had tumors in the rectum (10 cases, 22%), cecum, or anal canal (2 cases, 4% for each location, Table 1). Adenocarcinoma (AC) was the most common subtype (49/50, 98%, Table 1).

### Overlap of white blood cell-derived mutations (WDMs) with plasma-derived mutations in CRC patients

It has been reported that plasma cell-free DNA (cfDNA) contains mutations not only from cancer cells, but also from noncancerous blood cell precursors that harbor germline mutations or

**Table 1.** Summary of patients' clinical characteristics.

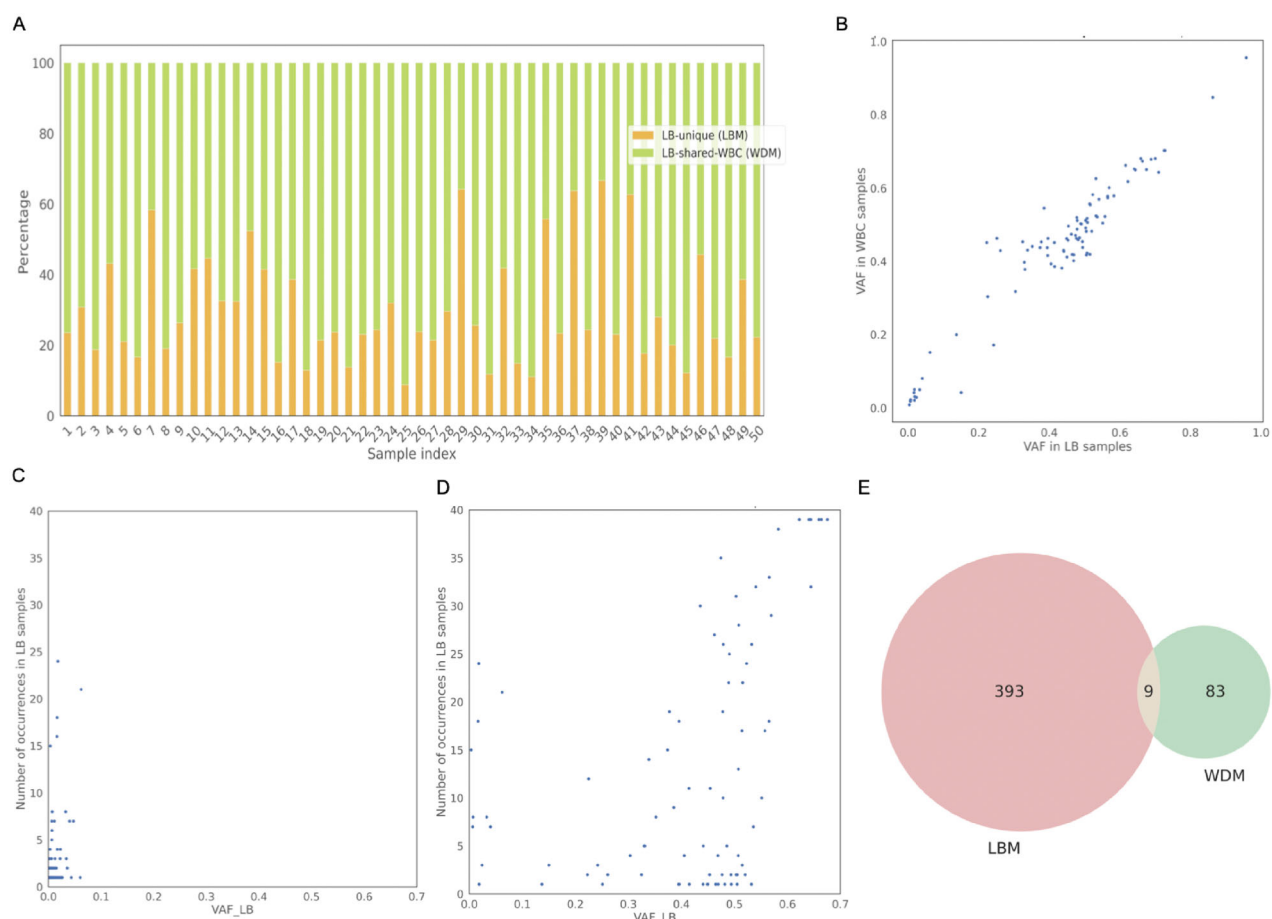
Patients' characteristics	Cancer (N=50) N (%)	Healthy controls (N=96) N (%)	$p$ value
Gender			
Female	18 (36)	58 (63.7)	<0.001
Male	32 (64)	28 (30.8)	
Unknown	0 (0)	5 (5.5)	
Age			
Median	63	34	<0.001
Min-Max	27–85	24–71	
Histology			
Adenocarcinoma	49 (98)		
Tubular adenomas	1 (2)		
Tumor Location			
Colon	35 (70)		
Rectum	11 (22)		
Cecum	2 (4)		
Anal canal	2 (4)		
Tumor stage			
0	2 (4)		
I	5 (10)		
II	20 (40)		
III	21 (42)		
Unknown	2 (4)		

N: total case number.

The Mann-Whitney test was conducted to compare the median age of patients at diagnosis. Chi-squared ( $\chi^2$ ) test was performed to compare gender and mutation frequencies.

mutations from clonal hematopoiesis of indeterminate potential (CHIP) (16,29). To distinguish such mutations from potential cancer-derived mutations, we parallely sequenced plasma cfDNA and paired white blood cell (WBC) genomic DNA, from all 50 CRC patients. Our sequencing assay examined the 20 most frequently mutated genes in CRC, according to the COSMIC database (Table S1) (22), and we also employed UMI technology to suppress sequencing error (20). DNA sequencing data was obtained from all 50 CRC patients with comparable on-target rates (62.5% and 67.5% for plasma cfDNA and WBC gDNA, respectively,  $p > 0.5$ ), and UMI consensus read coverage  $\geq 200X$ , for all types of samples (Figure S1(B) and (C)).

Specifically, we found that most mutations in the LB samples greatly overlapped with those from matched WBCs, across the 50 patients (median 75.9%; range: 33.3%–91.2%, Figure 1(A), Table 2), and the abundance of those mutations in plasma highly correlated with their abundance in WBCs ( $r = 0.95$ , 95% CI 0.94–0.97,  $p < 0.0001$ , Figure 1(B)). This finding further confirmed that WBC-derived mutations (WDMs) are the major constituents of CRC patient LBs and are unlikely to be sequencing errors or artifacts. WDMs were detected in 18 of the 20 selected genes, with



**Figure 1.** Detection of white blood cell-derived mutations by paired sequencing of white blood cell (WBC) gDNA and liquid biopsy (LB) cfDNA. (A) Detection rates of mutations shared between liquid biopsies and paired WBCs (WDMs) and mutations uniquely found in liquid biopsy samples (LBM) ( $n = 42$ ). (B) Correlation of the mean VAFs of WDMs in WBCs and LBs.  $p$ -values and correlation coefficients ( $r$ ) were calculated using Pearson's correlation test. (C,D) Scatter plots showing VAFs and occurrences of TDMs (C) and WDMs (D). (E) Venn diagram showing overlapping TDM and WDM spectra.

**Table 2.** Summary of mutation fractions in liquid biopsies and tumor tissue biopsies.

	WDM (%)	LBM (%)	FFPE-unique (%)	FFPE-shared-LB (TDM) (%)	FFPE-shared-WBC-not-LB (TIL) (%)	FFPE-shared-WBC-and-LB (WDM-FFPE) (%)
Mean	69.9	30.1	54.9	3.6	3.4	38.1
Std	15.6	15.6	22.5	4.1	2.5	20.3
Min	33.3	8.8	9.1	0.0	0.0	6.9
25%	59.2	19.3	43.7	1.2	1.9	24.3
50%	75.9	24.1	58.1	2.5	3.1	34.2
75%	80.7	40.8	69.5	4.5	4.4	49.4
Max	91.2	66.7	89.0	23.1	12.9	85.7

WDM: white blood cell (WBC) mutations overlapping with mutations in paired liquid biopsies (LB).

LBM: LB mutations after excluding WDM.

FFPE-unique: mutations in FFPE tissue samples not sharing with either WBC derived mutations or LBM.

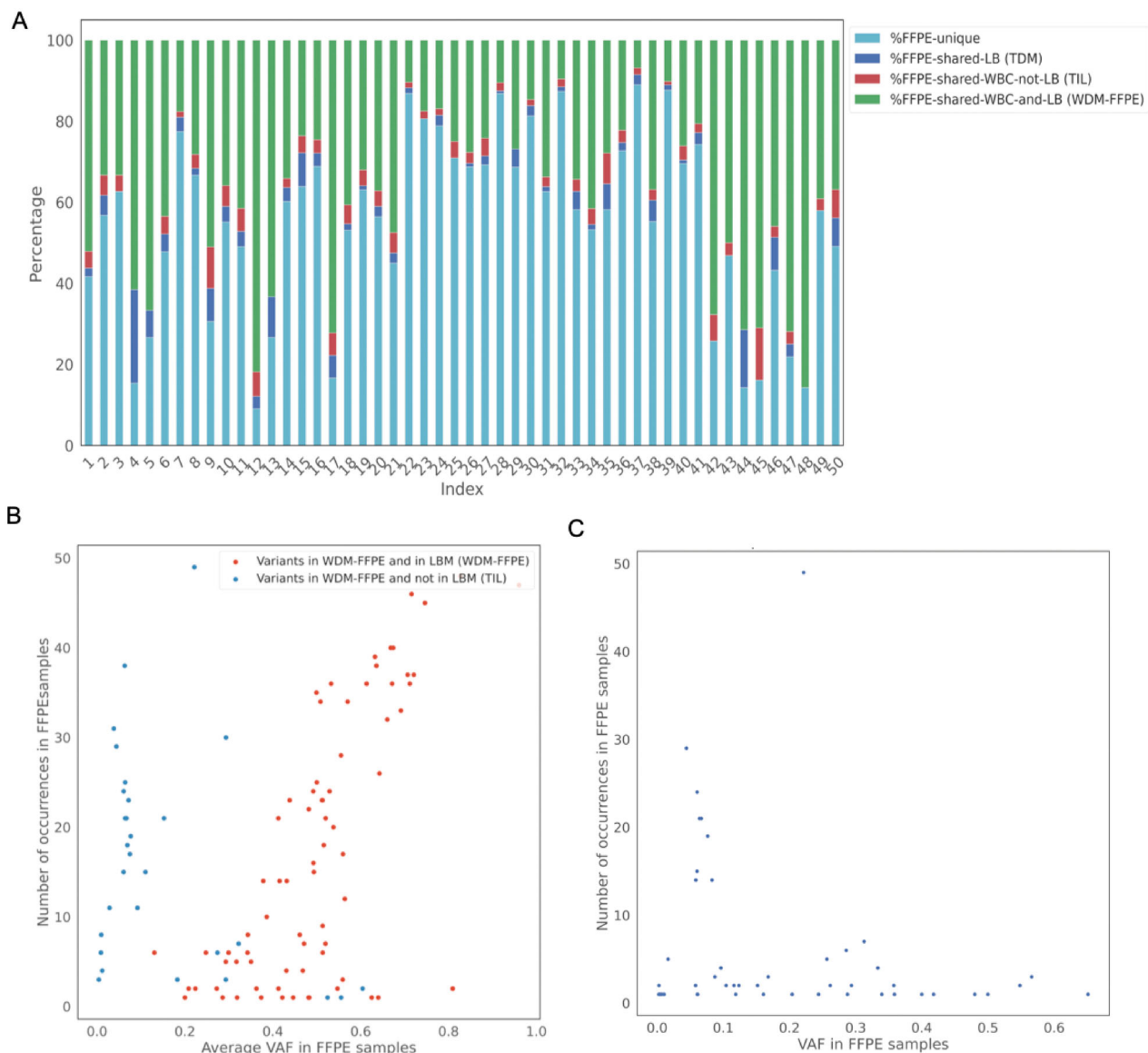
FFPE-shared-LB: LBM concordantly detected in paired tumor tissues, also defined as tumor derived mutations (TDM).

FFPE-shared-WBC-not-LB: mutations in FFPE overlapping with WBC mutations but not with LBM, also considered as tumor infiltrating lymphocytes (TIL).

FFPE-shared-WBC-LB: mutations in FFPE overlapping with both WBC and LBM mutations, also defined as WDM-FFPE.

*KMT2C* being the most frequently mutated gene (22.8% of all WDMs, Figure S2). After excluding WDMs, the remaining non-overlapping mutations, denoted as LB-unique mutations (LBMs), accounted for lower proportions of total LB-derived mutations, but being potentially of tumor

origin (median: 24.1%, range: 8.8%–66.7%, Figure 1(A), Table 2). While LBMs were mostly present as low occurrence mutations and variant allele frequencies (VAFs)  $< 0.1$  (Figure 1(C)), WDMs had varying occurrences (range: 1–39 patients, Figure 1(D)) and VAFs (range: 0.001–0.67, Figure



**Figure 2.** Identification of tumor-derived mutations (TDMs) in plasma, by sequencing paired tumor tissues. (A) Proportions of different mutation groups in tumor tissues from 50 patients, including mutations overlapping WBC mutations (WDM-FFPE and TIL), with LBMs (TDMs) or those uniquely detected in tumor tissues (unique-FFPE). (B) Scatter plot showing the occurrences (y axis) and mean VAFs (x axis) of WDM-FFPE (blue circles) and TILs (red circles). (C) Scatter plot showing occurrences (y axis) and VAFs (x axis) of TDM concordantly detected in LB and paired FFPE.

1(D)). Among detected WDMs, the majority were detected at VAFs  $>0.2$ , indicating that they were mostly derived from germline mutations (12), while the remainders had VAFs  $<0.1$ , comparable to the VAFs of LBMs (VAF  $< 0.1$ ), which could be CHIP related mutations (Figure 1(C)). We next cross-compared the profiles of WDMs and LBMs, among 50 patients, to examine if they could distinctly classify one from another. Of the total 92 detected WDMs, 9 overlapped with LBMs detected across individual patients, indicating that the spectrums of WDMs

and LBMs were not distinct, or that a WDM in a particular patient could be a TDM in another (Figure 1(E)). Together, these data showed that paired sequencing of WBC gDNA and plasma cfDNA may be required to distinguish WDMs from LBMs, in liquid biopsy samples.

#### **Identification of tumor-derived mutations in liquid biopsies by paired sequencing of tumor tissues**

To identify possible tumor origin of LBMs, we performed sequencing on patient-paired tumor

**Table 3.** Comparison of occurrences of TDMs shared between cancer and healthy controls.

	Variant	Control (N = 96)		Cancer (N = 50)		chi2_p_value, Bonferroni corrected
		Occurrences	Frequency (%)	occurrences	Frequency (%)	
1	chr17:7670685-G>A	1	1.0	1	2.0	0.78
2	chr12:25245350-C>A	1	1.0	1	2.0	0.78
3	chr17:7674220-C>T	1	1.0	1	2.0	0.78
4	<b>chr12:49033896-C&gt;T</b>	<b>1</b>	<b>1.0</b>	<b>15</b>	<b>30.0</b>	<b>&lt;0.0001 0.0000149</b>
5	chr1:26731393-C>G	1	1.04	2	4.0	0.56
6	chr7:152235912-C>T	15	15.6	7	14.0	0.99
7	chr7:152265049-G>T	12	12.5	6	12.0	0.86
8	chr5:112839942-C>T	3	3.1	7	14.0	0.03
9	chr7:152248171-G>A	32	33.3	24	48.0	0.12
10	chr7:152235831-C>G	1	1.0	1	2.0	0.78
11	chr4:186596870-C>T	1	1.0	1	2.0	0.78
12	chr7:152235905-C>T	14	14.6	7	14.0	0.88
13	chr7:152248016-G>C	81	84.4	41	82.0	0.89
14	chr12:25225628-C>T	1	1.0	1	2.0	0.78
15	chr7:152273712-A>T	33	34.4	18	36.0	0.99

Chi-squared ( $\chi^2$ ) test was performed to compare mutation frequencies between the CRC and control cohorts, and the *p*-values were subsequently adjusted by Bonferroni correction.

tissues. Like LB samples, we detected mutations in tumor tissues that overlapped with paired WBC mutations, across 50 CRC patients (Figure 2(A)). Those WBC shared mutations formed 2 distinct clusters with the first group overlapping with plasma WDM (WDM-FFPE) and the remaining group not shared with WDM (Figure 2(A)). WDM-FFPE accounted for a median of 34.2% of all mutations detected in all 50 paired patient tumors (range: 6.9%–85.7%, Figure 2(A), Table 2) and had relatively high VAFs >0.2 (Figure 2(B), red circles). By contrast, the remaining group accounted for lower proportions, with a median of 3.1% (range: 0%–12.9%) in 46/50 (92%) patients (Figure 2(A), Table 2) and showed lower VAFs than WDM-FFPE, possibly representing those derived from tumor-infiltrating lymphocytes (TILs) (Figure 2(B), blue circles). These data suggest that WBC-derived mutations are present not only in plasma, but also in CRC tumor tissues, which could interfere with the identification of cancer-specific mutations.

Strikingly, only a small proportion of mutations in tumor tissues overlapped with LBMs (median: 2.5%, range 0–23.1%, Figure 2(A), Table 2), but were not shared with WDM-FFPE or TIL mutations, in paired plasma samples; these were denoted as tumor-derived mutations (TDMs) (Figure 2(A)). These data suggested that not all mutations detected in tumor tissues were shed into the circulation. Additionally, TDMs were detected in 42/50 (84%) tissue samples, with a

wide range of VAFs (range: 0.001–0.67, Figure 2(C)) indicating that TDMs could be originated from either common or minor tumor clones.

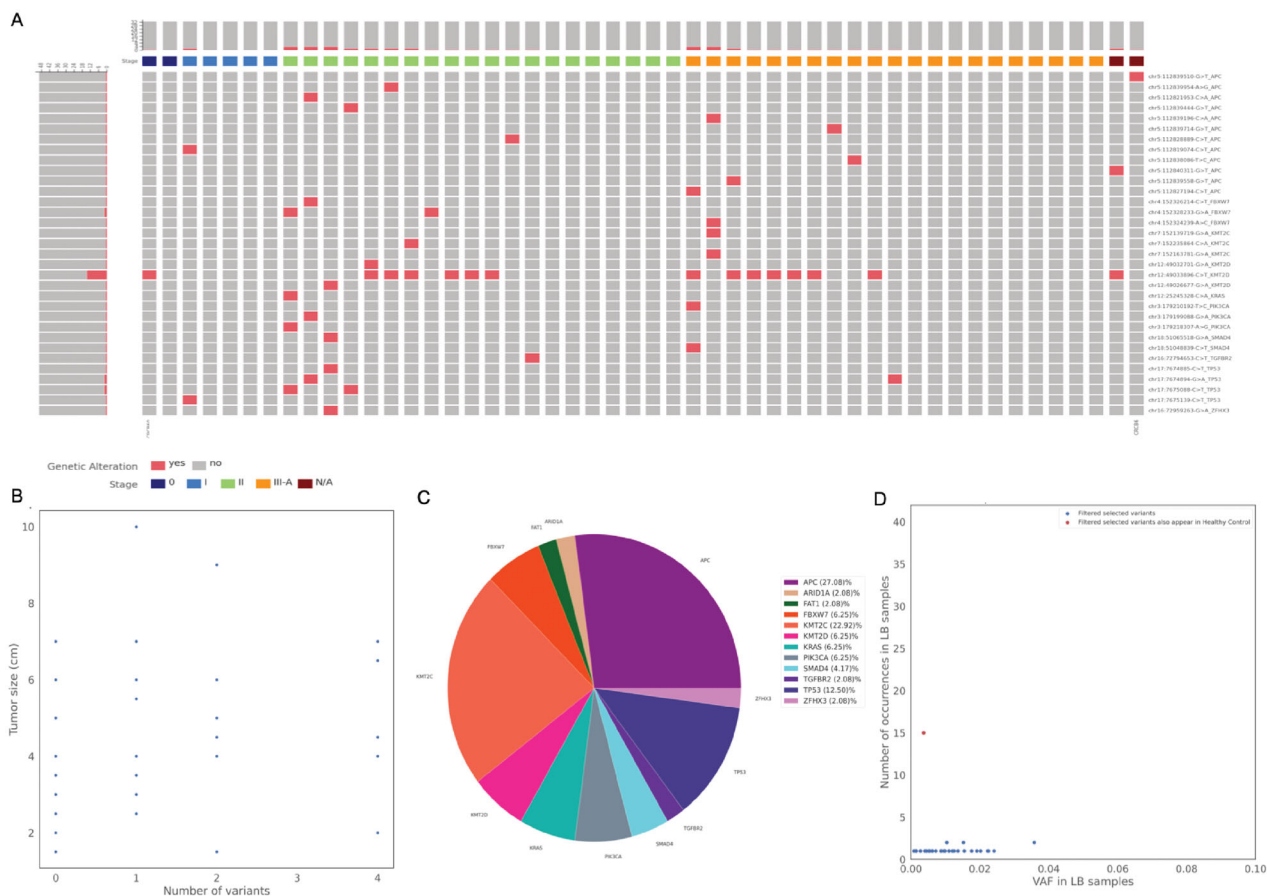
#### Overlap of TDMs with mutations detected in healthy individuals

It has been reported that background mutations including benign somatic mutations or mutations overlapping with WBC-derived mutations in healthy controls could lead to the false positive detection of TDM (12,18,30). To detect such overlapping mutations, we compared our TDM profile with that of the 96 healthy controls, finding that 16/48 (33.3%) overlapped. Of those shared mutations, one TDM was detected at a frequency significantly higher in cancer patients than in the control cohort (15/50, 30% in cancer patients versus 1/96 (1%) in healthy controls,  $p < 0.0001$ , Table 3), indicating high likelihood of being a true TDM. By contrast, the majority (15/16) of shared TDMs had frequencies comparable between cancer and healthy cohorts, making it unlikely to distinguish those background mutations from true TDMs (Table 3).

#### Characterization of TDMs in liquid biopsy samples of patients with nonmetastatic CRC

High specificity is critically important to minimize false positive results during early cancer diagnosis thereby preventing unnecessary follow-up tests and overtreatment (31). To achieve a high





**Figure 3.** Characteristics of tumor-derived mutations detected in liquid biopsy samples of CRC patients. (A) Oncoprint of distribution of TDMs in CRC patients, as correlated with different CRC stages. Rows and columns represent TDMs and patients, respectively. Mutations are labelled on the right side. The left-side bar plot represents the mutational loads of each patient. Cases are grouped into stages. Patients are grouped according to their tumor stages. (B) Correlation between number of TDMs and tumor size. Spearman's rank correlation coefficient was performed to analyze the correlations between tumor size and the number of detected TDMs in 28 CRC samples; ns: not significant. (C) Pie-chart showing the distribution of genes harboring TDMs. (D) Scatter plot showing occurrences and VAFs of TDMs in liquid biopsies.

specificity of TDM identification, we excluded potential nontumor-derived mutations by removing TDMs shared with healthy controls, but not significantly more frequent than the control cohort. After applying the above filtering method, we obtained a set of 33 TDMs, with 28/50 (56%) CRC patients having at least one (Figure 3(A)). Of those, stage II patients had the highest rate of TDM detection (13/20, 65%), while 10/21 (47.6%) were stage IIIA, and 2/7 (28.6%) stage 0-I (Figure 3(A)). In addition, the number of TDMs detected in LB samples varied across patients, with stage II patients having higher TDM loads than other stages, thus demonstrating intratumor heterogeneity (Figure 3(A)). There were no significant

correlations between the number of detected TDMs and tumor size (Figure 3(B)). Among the 33 detected TDMs, 18 (54.5%) were previously confirmed as pathogenic or likely pathogenic mutations in the ClinVar database (32), while the remaining had unknown functions, or were previously classified as benign (Table S3).

TDMs were most frequent in *APC* (36.4%), followed by *TP53* (12.1%), *KMT2C*, *FBXW7*, *KMT2D*, *PIK3CA* (9.1% for each gene), *SMAD4* (6.1%), *KRAS*, *TGFBR2*, and *ZFH33* (3% for each gene) (Figure 3C). The majority (29/33, 87.9%) of TDMs were present at low VAFs (<0.04), in liquid biopsies, and were not shared among CRC patients, thus highlighting the inter-individual

heterogeneity of TDMs (Figure 3(D)). Interestingly, one TDM, mapped to *KMT2D* (chr12: 49033896-C>T *KMT2D*), was recurrently detected in CRC patients, at a frequency of 30% (15/50) (Figure 3(D)).

## Discussion

In this study, we examined the feasibility of utilizing tumor-derived mutations, in ctDNA, for the purpose of early detection of colorectal cancer (CRC). While metastatic disease correlates with numerous genetic alterations, early-stage malignancies contain much fewer (33). Other challenges to identifying diagnostic ctDNA tumor-derived mutations (TDMs) include interference by white blood cell-derived mutations (i.e., WDMs, occurring during clonal hematopoiesis of indeterminate potential (CHIP)) (29), in addition to intra- and inter-tumor heterogeneity, benign somatic mutations, and errors inherent to the limitations of ultradeep sequencing (10,18,33).

To study such feasibility, we procured a distinctive, valuable patient set of 50 nonmetastatic (stage 0-I, II and IIIA) tumor tissues (FFPE), matched to white blood cell (WBC) and plasma specimens. The high proportion of CRC patients with stage IIIA (21/50, 42%) truly reflected the clinical context that late-stage cancer diagnosis remains dominant in Vietnam (34). However, in this instance, CRC patients with stage IIIA were known to have nonmetastatic disease, and cancer detection at this stage would provide high treatment success rates (35). To reduce errors associated with massive parallel sequencing, we added unique molecular identifiers (UMI) to sequenced fragments, an approach that has been shown to reduce such errors by >70-fold, with sensitivity up to 98% (36).

Consistent with previous findings, we found much higher proportions of WBC-derived mutations (WDMs) in plasma samples of CRC patients (16,37). WDMs detected in plasma samples could be a mixture of germline mutations or CHIP-related mutations from WBCs. While the former group can be identified as germline mutations, based on their high (>0.1) VAF levels, CHIP-related mutations not only showed comparable VAF levels, but their spectrums

overlapped with TDM spectra. These findings suggest that paired sequencing of WBC gDNA is required to differentiate TDMs from CHIP-related mutations. Consistent with a recent study by Chan et al. (37), we also detected a group of WBC-derived mutations in tumor tissues that did not overlap with tumor- or WBC-derived mutations detected in paired plasma samples, possibly derived from tumor-infiltrating lymphocytes (TILs), which have been reported to be involved in tumor immunosuppression (38). Strikingly, those TIL mutations, and TDMs, showed comparable VAFs in tumor tissues, which could lead to the inaccurate identification of TDMs (Figure 2(B,C)). Thus, further studies are required to dissect the possible implications of TIL mutations to tumorigenesis. To accurately classify mutations detected in LB (LBMs) as tumor-derived, we excluded TIL mutations, and selected only LBMs concordantly present in paired tumor tissue samples. The remaining nonoverlapping LBMs could be derived either from tumor clones that were lost during tissue sampling or from unknown sources, as previously described by Razavi et al. (16).

Background mutations from healthy subjects could lead to the misinterpretation of TDMs (12,18,30). However, this challenge has not been well addressed in previous studies. Here, we profiled mutations in plasma samples of 96 healthy control subjects who underwent colonoscopy but had no lesions. We detected the overlapping spectrum between TDMs and mutations from healthy subjects, confounding the distinction of cancerous mutations from background mutations. After filtering potential background mutations, we detected 33 TDMs, highly likely of tumor origin, in 56% (28/50) of CRC patients, comparable to the detection rates of ctDNA mutations reported by previous studies for early CRC stages (5,12). Moreover, like others (5,12), we also found that higher-stage disease correlated with greater numbers of tumor-matching ctDNA-unique mutations. Those TDMs were mapped to 10 of 20 CRC genes, with *APC* and *TP53* accounting for those most frequently mutated. The majority of identified TDMs had low occurrences, demonstrating the interindividual heterogeneity of early-stage tumors. Noticeably, we

identified one TDM, a silent mutation mapped to *KMT2D*, that was recurrently detected in 15 CRC patients. Despite being a nonsynonymous mutation, this TDM is located at the end of exon 40 of *KMT2D*, suggesting that it possibly alters splicing regulatory sites, mRNA stability and miRNA binding sites, thus might be related to tumorigenesis (39). Thus, future studies, using larger sample sizes, might be required to confirm the tumor origin, as well as the functional role(s) of this potential “hotspot” TDM.

There are several limitations to our study. First, the use of paired tumor tissues to confirm the tumor origin of TDMs could result in loss of TDMs shed by tumor clones that were not captured by tissue sampling (8,40). Second, paired sequencing of WBCs and LBs, for the 96 healthy controls, was not carried out in the current study to identify CHIP mutations, due to cost constraints. By filtering mutations shared with healthy subjects, we might have excluded some true TDMs that overlapped with low-frequency CHIP mutations from WBCs from healthy subjects. However, this method will allow us to achieve the best specificity of TDM detection, and avoid overdiagnosis (i.e., false positives), which is an important criterion for early cancer detection. A probabilistic model may be required to accurately classify TDMs from background mutations in future studies. Third, the healthy control cohort in this study had a median age younger than the cancer cohort, which may have an impact on the identified mutational profiles, particularly CHIP-related mutations. Previous studies have reported a link between individual’s ages and detection rates of CHIP-associated mutations (41,42). In addition, our study lacks clinical follow-up with information on the health and disease status of the healthy subjects. This is important since a healthy individual may carry cancer-related mutations and subsequently develop cancer (43). Hence, the present study is a cohort study, and future case-control studies with larger data sets and age ranges better balanced between control and cancer patients, together with robust statistical models, are required to improve the accuracy of TDM detection.

In summary, we set forth a workflow for identifying tumor-specifically-derived circulating

DNA mutations in early-stage colorectal cancer. However, our results indicated that most TDMs showed both intratumoral and intertumoral heterogeneity, resulting in insufficient sensitivity for widespread CRC detection. Thus, we assert that combination with other ctDNA biomarkers such as methylated DNA, varying fragmentation patterns, and altered chromosomal copy numbers, could increase the positive predictive value of liquid biopsies, and warrant more in-depth study of these phenomena.

### Acknowledgement

The author would like to thank Dr. Curt Balch (Bioscience Advising, Ann Arbor, MI, USA) for proofreading the manuscript.

### Author contributions

Conceptualization, Anh Vu Hoang, Hoang Bac Nguyen, Kiet Truong Dinh, Hoai-Nghia Nguyen, Thanh-Thuy Thi Do, Hoa Giang and Diep Tuan Tran; Data curation, Huu-Thinh Nguyen, Duc-Huy Tran, Quoc Dat Ngo, Truong Vinh Ngoc Pham, Minh Triet Le, Ngoc An Trinh Le, Trung Kien Le and Le Son Tran; Formal analysis, Bac An Luong, Le Gia Hoang Le, Quoc Chuong Ho, Hue-Hanh Thi Nguyen, Cao Minh Nguyen, Vu Uyen Tran, Thanh Luan Nguyen, Hong-Anh Thi Pham, Hong Thuy Le, Hong Diep Thi Duong and Le Son Tran; Funding acquisition, Thanh-Thuy Thi Do and Diep Tuan Tran; Methodology, Huu-Thinh Nguyen, Hoa Giang and Le Son Tran; Project administration, Huu-Thinh Nguyen, Hoai-Nghia Nguyen and Le Son Tran; Software, Trong-Hieu Nguyen, Minh-Duy Phan and Hoa Giang; Supervision, Diep Tuan Tran; Writing – original draft, Le Son Tran; Writing – review & editing, Hoai-Nghia Nguyen and Diep Tuan Tran.

### Institutional review board statement

this study was approved by The Ethics Committee of University of Medicine and Pharmacy at Ho Chi Minh City, Vietnam (Ethic number: 383/ĐHYD-HĐĐĐ)

### Informed consent statement

Informed consent was obtained from all subjects involved in the study.

### Declaration of interest

No potential conflict of interest was reported by the author(s).

## Funding

This research was funded by Ho Chi Minh city Department of Science and Technology under grant number 52/2019/HĐ-QP T KHCN (to DT T).]

## ORCID

Minh-Duy Phan  <http://orcid.org/0000-0002-3426-1044>

Le Son Tran  <http://orcid.org/0000-0002-5382-3903>

## References

- Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. *CA A Cancer J Clin.* 2020;70(1):7–30. doi:10.3322/caac.21590.
- Issa IA, Nouredine M. Colorectal cancer screening: an updated review of the available options. *World J Gastroenterol.* 2017;23(28):5086–96. doi:10.3748/wjg.v23.i28.5086.
- Bressler B, Paszat LF, Chen Z, Rothwell DM, Vinden C, Rabeneck L. Rates of new or missed colorectal cancers after colonoscopy and their risk factors: a population-based analysis. *Gastroenterology.* 2007;132(1):96–102. doi:10.1053/j.gastro.2006.10.027.
- Rawla P, Sunkara T, Barsouk A. Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors. *Prz Gastroenterol.* 2019;14(2):89–103. doi:10.5114/pg.2018.81072.
- Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science.* 2018;359(6378):926–30. doi:10.1126/science.aar3247.
- Wan JCM, Massie C, Garcia-Corbacho J, Mouliere F, Brenton JD, Caldas C, et al. Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat Rev Cancer.* 2017;17(4):223–38. doi:10.1038/nrc.2017.7.
- Bettegowda C, Sausen M, Leary RJ, Kinde I, Wang Y, Agrawal N, et al. Detection of circulating tumor DNA in early- and late-stage human malignancies. *Sci Transl Med.* 2014;6(224):224ra24. doi:10.1126/scitranslmed.3007094.
- Nguyen HT, Tran DH, Ngo QD, Pham HT, Tran TT, Tran VU, et al. Evaluation of a liquid biopsy protocol using ultra-deep massive parallel sequencing for detecting and quantifying circulation tumor DNA in colorectal cancer patients. *Cancer Invest.* 2020;38(2):85–93. doi:10.1080/07357907.2020.1713350.
- Tran LS, Nguyen Q-TT, Nguyen CV, Tran V-U, Nguyen T-HT, Le HT, et al. Ultra-deep massive parallel sequencing of plasma cell-free dna enables large-scale profiling of driver mutations in Vietnamese patients with advanced non-small cell lung cancer. *Front Oncol.* 2020;10(1351):1351. doi:10.3389/fonc.2020.01351.
- Ma F, Guan Y, Yi Z, Chang L, Li Q, Chen S, et al. Assessing tumor heterogeneity using ctDNA to predict and monitor therapeutic response in metastatic breast cancer. *Int J Cancer.* 2020;146(5):1359–68. doi:10.1002/ijc.32536.
- Ignatiadis M, Sledge GW, Jeffrey SS. Liquid biopsy enters the clinic – implementation issues and future challenges. *Nat Rev Clin Oncol.* 2021;18(5):297–312. doi:10.1038/s41571-020-00457-x.
- Phallen J, Sausen M, Adleff V, Leal A, Hruban C, White J, et al. Direct detection of early-stage cancers using circulating tumor DNA. *Sci Transl Med.* 2017;9(403):2,7,8, 10. doi:10.1126/scitranslmed.aan2415.
- Abbosh C, Birkbak NJ, Wilson GA, Jamal-Hanjani M, Constantin T, Salari R, et al. Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature.* 2017;545(7655):446–51. doi:10.1038/nature22364.
- Strom SP. Current practices and guidelines for clinical next-generation sequencing oncology testing. *Cancer Biol Med.* 2016;13(1):3–11.
- Cristiano S, Leal A, Phallen J, Fiksel J, Adleff V, Bruhm DC, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature.* 2019;570(7761):385–9. doi:10.1038/s41586-019-1272-6.
- Razavi P, Li BT, Brown DN, Jung B, Hubbell E, Shen R, et al. High-intensity sequencing reveals the sources of plasma circulating cell-free DNA variants. *Nat Med.* 2019;25(12):1928–37. doi:10.1038/s41591-019-0652-7.
- Ma M, Zhu H, Zhang C, Sun X, Gao X, Chen G. Liquid biopsy"-ctDNA detection with great potential and challenges. *Ann Transl Med.* 2015;3(16):235.
- Fiala C, Diamandis EP. Mutations in normal tissues-some diagnostic and clinical implications. *BMC Med.* 2020;18(1):283. doi:10.1186/s12916-020-01763-y.
- Tian G, Xia L, Li Z, Li X, Xu F, Liu C, He J. Baseline mutation profiling of circulating cell-free DNA from healthy individuals to improve the detection accuracy of circulating tumor DNA in cancers. *JCO.* 2017;35(15\_suppl):e23057–e23057. doi:10.1200/JCO.2017.35.15\_suppl.e23057.
- Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res.* 2017;27(3):491–9. doi:10.1101/gr.209601.116.
- Chen W, Li Y, Easton J, Finkelstein D, Wu G, Chen X. UMI-count modeling and differential expression analysis for single-cell RNA sequencing. *Genome Biol.* 2018;19(1):70. doi:10.1186/s13059-018-1438-9.
- Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br J Cancer.* 2004;91(2):355–8. doi:10.1038/sj.bjc.6601894.

23. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114–20. doi:[10.1093/bioinformatics/btu170](https://doi.org/10.1093/bioinformatics/btu170).
24. Li H, Durbin R. Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics*. 2010;26(5):589–95. doi:[10.1093/bioinformatics/btp698](https://doi.org/10.1093/bioinformatics/btp698).
25. <https://github.com/fulcrumgenomics/fgbio>. 2019. [updated 2019; cited 2021 5th July 2021].
26. <https://github.com/fulcrumgenomics/fgbio/blob/master/src/main/scala/com/fulcrumgenomics/umi/CallMolecularConsensusReads.scala>. 2019. [updated 2019; cited 2021 5th July 2021].
27. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res*. 2016;44(11):e108. doi:[10.1093/nar/gkw227](https://doi.org/10.1093/nar/gkw227).
28. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The ensembl variant effect predictor. *Genome Biol*. 2016;17(1):122. doi:[10.1186/s13059-016-0974-4](https://doi.org/10.1186/s13059-016-0974-4).
29. Boettcher S, Ebert BL. Clonal hematopoiesis of indeterminate potential. *J Clin Oncol*. 2019;37(5):419–22. doi:[10.1200/JCO.2018.79.3588](https://doi.org/10.1200/JCO.2018.79.3588).
30. Liu J, Chen X, Wang J, Zhou S, Wang CL, Ye MZ, et al. Biological background of the genomic variations of cf-DNA in healthy individuals. *Ann Oncol*. 2019;30(3):464–70. doi:[10.1093/annonc/mdy513](https://doi.org/10.1093/annonc/mdy513).
31. Liu MC, Oxnard GR, Klein EA, Swanton C, Seiden MV. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Ann Oncol*. 2020;31(6):745–59. doi:[10.1016/j.annonc.2020.02.011](https://doi.org/10.1016/j.annonc.2020.02.011).
32. Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, Hart J, et al. ClinVar: improvements to accessing data. *Nucleic Acids Res*. 2020;48(D1):D835–d44. doi:[10.1093/nar/gkz972](https://doi.org/10.1093/nar/gkz972).
33. Rossi G, Ignatiadis M. Promises and pitfalls of using liquid biopsy for precision medicine. *Cancer Res*. 2019;79(11):2798–804. doi:[10.1158/0008-5472.CAN-18-3402](https://doi.org/10.1158/0008-5472.CAN-18-3402).
34. Pham T, Bui L, Kim G, Hoang D, Tran T, Hoang M. Cancers in Vietnam—Burden and control efforts: a narrative scoping review. *Cancer Control*. 2019;26(1):1073274819863802. doi:[10.1177/1073274819863802](https://doi.org/10.1177/1073274819863802).
35. Edge SB, Compton CC. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. *Ann Surg Oncol*. 2010;17(6):1471–4. doi:[10.1245/s10434-010-0985-4](https://doi.org/10.1245/s10434-010-0985-4).
36. Forshew T, Murtaza M, Parkinson C, Gale D, Tsui DW, Kaper F, et al. Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci Transl Med*. 2012;4(136):136ra68.
37. Chan HT, Nagayama S, Chin YM, Otaki M, Hayashi R, Kiyotani K, et al. Clinical significance of clonal hematopoiesis in the interpretation of blood liquid biopsy. *Mol Oncol*. 2020;14(8):1719–30. doi:[10.1002/1878-0261.12727](https://doi.org/10.1002/1878-0261.12727).
38. Chiou SH, Sheu BC, Chang WC, Huang SC, Hong-Nerng H. Current concepts of tumor-infiltrating lymphocytes in human malignancies. *J Reprod Immunol*. 2005;67(1–2):35–50. doi:[10.1016/j.jri.2005.06.002](https://doi.org/10.1016/j.jri.2005.06.002).
39. Sharma Y, Miladi M, Dukare S, Boulay K, Caudron-Herger M, Groß M, et al. A pan-cancer analysis of synonymous mutations. *Nat Commun*. 2019;10(1):2569. doi:[10.1038/s41467-019-10489-2](https://doi.org/10.1038/s41467-019-10489-2).
40. Tran LS, Pham HT, Tran VU, Tran TT, Dang AH, Le DT, et al. Ultra-deep massively parallel sequencing with unique molecular identifier tagging achieves comparable performance to droplet digital PCR for detection and quantification of circulating tumor DNA from lung cancer patients. *PLOS One*. 2019;14(12):e0226193. doi:[10.1371/journal.pone.0226193](https://doi.org/10.1371/journal.pone.0226193).
41. Wu H-T, Kalashnikova E, Mehta S, Salari R, Sethi H, Zimmermann B, et al. Characterization of clonal hematopoiesis of indeterminate potential mutations from germline whole exome sequencing data. *JCO*. 2020;38(15\_suppl):1525. doi:[10.1200/JCO.2020.38.15\\_suppl.1525](https://doi.org/10.1200/JCO.2020.38.15_suppl.1525).
42. Jaiswal S, Ebert BL. Clonal hematopoiesis in human aging and disease. *Science*. 2019;366(6465):eaan4673. doi:[10.1126/science.aan4673](https://doi.org/10.1126/science.aan4673).
43. Gormally E, Vineis P, Matullo G, Veglia F, Caboux E, Le Roux E, et al. TP53 and KRAS2 mutations in plasma DNA of healthy subjects and subsequent cancer occurrence: a prospective study. *Cancer Res*. 2006;66(13):6871–6. doi:[10.1158/0008-5472.CAN-05-4556](https://doi.org/10.1158/0008-5472.CAN-05-4556).