

RESEARCH ARTICLE OPEN ACCESS

Tumor Genomic and Transcriptomic Analysis Integrated With Liquid Biopsy ctDNA Monitoring: Analytical Validation and Clinical Insights

Nam H. B. Tran^{1,2} | Thien-Phuc Hoang Nguyen^{1,2} | Vinh Quang Bui³ | Vu Thuong Le⁴ | Trong Khoa Mai^{5,6} | Van-Anh Nguyen Hoang^{1,2} | Tien Anh Nguyen^{1,2} | Minh-Duc Nguyen^{1,2} | Ha-Hieu Pham² | Tho Thi Le Vo^{1,2} | My T. T. Ngo^{1,2} | Du Quyen Nguyen^{1,2} | Duy Sinh Nguyen^{1,2} | Hoai-Nghia Nguyen^{1,2} | Minh-Duy Phan^{1,2} | Hoa Giang^{1,2} | Lan N. Tu^{1,2} 

¹Medical Genetics Institute, Ho Chi Minh City, Vietnam | ²Gene Solutions, Ho Chi Minh City, Vietnam | ³Hanoi Oncology Hospital, Hanoi, Vietnam | ⁴University Medical Center of HCMC, Ho Chi Minh City, Vietnam | ⁵Bach Mai Hospital, Hanoi, Vietnam | ⁶Ha Noi Medical University, Hanoi, Vietnam

Correspondence: Lan N. Tu (lantu@genesolutions.vn)

Received: 15 September 2025 | **Revised:** 19 November 2025 | **Accepted:** 20 November 2025

Keywords: cancer of unknown primary | cancer tissue of origin | circulating tumor DNA | comprehensive genomic profiling | copy number variation | fusion | mRNA sequencing

ABSTRACT

Background: Comprehensive genomic profiling (CGP) is a time- and tissue- efficient method to help guide precision oncology. To enhance the clinical utility of CGP, we investigated the performance of a novel strategy integrating tumor DNA and mRNA profiling, together with liquid biopsy ctDNA monitoring.

Methods: Genomic DNA and mRNA simultaneously extracted from 604 archived tissue samples of 12 cancer types were used. Tumor DNA was subjected to targeted sequencing using a 504-gene panel with high-density probes (HDP), and shallow whole genome sequencing to profile genomic biomarkers. mRNA transcriptome profiling was performed to further capture fusion variants, and to predict tissue of origin (TOO) using our ensemble model OriCUP, an algorithm trained on 9889 samples and independently validated on 731 samples. In a cohort of 55 metastatic lung cancer patients, longitudinal plasma ctDNA was analyzed using a hybrid tumor-informed and tumor-agnostic approach to predict progression-free survival (PFS).

Results: Among all biomarkers, DNA sequencing using HDP achieved higher sensitivity than the standard panel design to identify copy number variations at chromosome-, gene-, and exon- levels. The detection rate of fusion variants using DNA sequencing alone was 20% lower than mRNA sequencing in reference samples, while the combination of both methods was essential to maximize fusion detection in clinical FFPE samples. For TOO, our OriCup model achieved prediction accuracy of 87.7% for primary tumors and 81.4% for metastatic tumors. In 55 lung cancer patients, ctDNA profiling identified additional 11.5% tumor-agnostic actionable and resistance mutations. Patients having more than 50% ctDNA decrease from baseline were classified as molecular responders and showed significantly longer PFS than those classified as molecular non-responders (HR = 9.42, 95% CI: 3.33–26.67, $p < 0.0001$, 12-month PFS: 95.5% vs. 31.7%).

Conclusions: Comprehensive genomic and transcriptomic profiling could reliably unveil genetic details not provided by DNA-only CGP. The integration of ctDNA detection further helped detect tumor-agnostic mutations and monitor treatment response.

Nam H. B. Tran and Thien-Phuc Hoang Nguyen contributed equally to this work.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). *Cancer Medicine* published by John Wiley & Sons Ltd.

1 | Introduction

The fast pace of biomarker discovery and increasing number of approved targeted therapies are driving the demand for comprehensive genomic profiling (CGP) in precision oncology [1]. Unlike single-gene assays or small gene panels, a comprehensive approach maximizes the chances of identifying relevant biomarkers for patient stratification, predicting drug response, and monitoring resistance mechanisms. The majority of current CGP utilizes formalin-fixed paraffin-embedded (FFPE) tissue samples as their high tumor content allows for high sensitivity and specificity in detecting genetic alterations [2]. FFPE DNA is more commonly used than FFPE mRNA because DNA, despite damages from the formalin fixation process, is generally more stable and robust compared to mRNA. However, mRNA profiling is particularly sensitive to assess novel or atypical gene fusions that DNA analysis alone may not detect [3]. Moreover, the tumor transcriptome could help predict the tissue of origin (TOO) for cancer of unknown primary (CUP), which can then be combined with the genetic alteration profile to guide more effective site-specific treatment for patients with CUP [4]. Although it is evident that mRNA profiling could complement DNA information to expand therapeutic options, the current practice of running mRNA and DNA testing separately is inefficient in terms of both time and tissue utilization.

While tissue biopsies provide a direct snapshot of the molecular landscape of primary tumors, they are invasive, subject to sampling bias, and may not fully capture tumor heterogeneity or dynamic changes over time [5, 6]. Liquid biopsy (LB), conversely, is minimally invasive and serial monitoring of circulating tumor DNA (ctDNA) in LB assists real-time tumor burden assessment and hence prediction of treatment response [6]. LB-ctDNA profiling is typically performed as a separate test from FFPE profiling, resulting in less comprehensive information and ultimately higher costs for patients.

Since FFPE DNA and mRNA profiling, together with LB-ctDNA detection, each offer unique and complementary information, combining them into a single affordable test could potentially unlock more treatment opportunities for cancer patients. In this study, we validated the technical performance and presented the clinical relevance of our unified approach that integrated the profiling of tumor DNA using high-density probes, tumor mRNA transcriptome, and LB-ctDNA.

2 | Materials and Methods

2.1 | Sample and Data Collection

A total of 604 archived tissue samples of 12 cancer types were obtained from the Medical Genetics Institute, Viet Nam. All samples were de-identified and stored for <2 years. For FFPE-derived biomarkers, 241 and 375 tissue samples were used for DNA and mRNA analysis respectively. Specifically for TOO prediction, transcriptome data of 9889 samples from TCGA database [7] and 356 samples from Robinson et al. [8] were added to the analysis. For LB-ctDNA assessment, 147 real-world plasma samples collected from a separate cohort of 55

stage-IV lung cancer patients were analyzed and compared with clinical records, following our published study design [9]. Detailed sample information for each analysis is listed in Table S1.

The study reported within this manuscript was approved by the Institutional Ethics Review Board of the Medical Genetics Institute: approval 03/2024/CT-VDTYH and 03/2023/CT-VDTYH. All participants provided written informed consent for the use of their anonymized genomic and clinical information. The study was performed in accordance with the Declaration of Helsinki and local guidelines/regulations.

2.2 | Sample Processing

Genomic DNA and total RNA were co-isolated from FFPE by AllPrep DNA/RNA FFPE Kit (Qiagen, Germany). DNA libraries from matched FFPE and white blood cells (WBC) were prepared as described previously [10] and hybridized with probes targeting 504 gene (Integrated DNA Technologies, USA) (Table S2). The standard panel design (STD) had equal probe density in all interrogated coding regions, while the high-density probe (HDP) panel was enriched with more probes in certain chromosomes and 27 genes that often have copy number variation (CNV). Total RNA was converted to cDNA library using NEBNext Ultra II Directional RNA Library Prep Kit (New England Biolabs, USA), followed by exome enrichment of 19,435 genes (xGen Exome Hyb Panel v2, Integrated DNA Technologies, USA). Enriched libraries were sequenced using DNBSEQ-T7 sequencer (MGI Tech, China). For shallow whole-genome sequencing (sWGS), FFPE DNA libraries were sequenced directly using DNBSEQ-G400 sequencer (MGI, China) with an average depth of 1-2X. Besides clinical samples, 14 reference samples (Table S3) were processed using the same protocol.

2.3 | FFPE DNA Sequencing Data Analysis

Somatic variant calling, using tumor-normal mode, and germline variant calling were performed using DRAGEN (v4.3) as described previously [10, 11]. CNV was analyzed using DRAGEN's somatic CNV pipeline (v4.3); gene amplification was defined when the ratio of segment mean was ≥ 2.0 , while gene loss was defined when the ratio was below 0.6. In silico tumor fraction (TF) was estimated using FACETS (v0.6.2), based on heterozygous SNPs present in both FFPE and WBC samples. The absolute copy number (ACN) was then estimated using the formula: $ACN = ([2 \times \text{ratio} - 2(1 - TF)] / (TF))$. For large genomic rearrangement (LGR) at exon-level, the DRAGEN Structural Variant Caller pipeline utilizing both paired-end and split-read mapping information was used. For homologous recombination deficiency (HRD) analysis, besides *BRCA* mutations, data from sWGS were processed to calculate the whole-genome instability (wGI) score as described previously [12]. Samples with a wGI score ≥ 0 were classified as positive for genomic instability. Chromosome-level structural variants (SVs) were also detected using sWGS data by calculating the log2 ratio of each target bin size relative to the median value across the genome, excluding sex chromosomes.

A log2 ratio below 0.5 indicated a loss event, while a log2 ratio above 1.3 indicated a gain event [13].

2.4 | FFPE mRNA Sequencing Data Analysis

For fusion detection, FASTQ files were pre-processed to remove low-quality bases before alignment to the GRCh38 reference genome. DRAGEN (v4.3) in RNA mode was used for read alignment and fusion detection. Samples were excluded from further analysis if they failed QC criteria: total input reads ≤ 30 million, median insert size < 100 bp, or coefficient of variation in median transcript coverage > 1 . A fusion event between 2 genes was considered positive if the number of junction-supporting reads was ≥ 5 . Fusions involving exon-skipping events within a single gene were considered positive if supported by more than 10 reads.

2.5 | Machine Learning Model for Tissue of Origin Prediction

Gene-level read counts from processed BAM files were obtained using the featureCounts tool from the Subread package (v.2.0.1) [14]. Raw read counts were normalized to remove batch effects using the TMM (Trimmed Mean of M -values) method in the edgeR package (v.4.1.2) and then converted to FPKM values. In addition to in-house samples, normalized expression profiles from 32 primary tumor types ($n=9889$) were downloaded from the UCSC Xena platform [7] in FPKM format (Table S1). These samples were split into training set (8–100 samples/cancer type) and testing set ($n=2803$ and 7086 samples, respectively) to optimize machine learning (ML) model. The independent validation dataset had 731 samples consisting of 355 metastatic tumors from Robinson et al. [8] and in-house 376 primary and metastatic tumors (Table S1).

For model training, we first established 3 gene sets: 90 genes from Ye et al. study [15]; 2000 high-variable genes selected by a modified SIMBA method [16] that identified top variable genes to distinguish cancers in the training set with gini score ≥ 0.3 ; and 739 high-variable genes that were further selected from the 2000 genes above to enable classification accuracy $> 50\%$ in common cancer types. Using the 3 gene sets, we then used the AutoML framework auto-sklearn (v.0.15.0), based on scikit-learn package, to train our OriCUP model with multi-class classification approach. The pipeline tested various supervised learning algorithms with hyperparameter optimization via 10-fold cross-validation. The final ensemble model was selected based on the lowest loss across validation folds, and the gene set was selected for the highest prediction performance in the validation dataset. For each sample, the top two predicted labels with the highest probabilities were reported.

The OriCUP model was compared with the deconvolution method [17] and the CUP-AI-Dx model [18]. Deconvolution was performed using FARDEEP (v.1.0.1), and normalized expression profiles from both TCGA cancer samples and GTEx normal tissues were used as input with 3 gene sets as signature markers for TOO. The CUP-AI-Dx model was downloaded from its GitHub repository (<https://github.com/TheJacksonLaboratory/>

CUP-AI-Dx) and the normalized TPM data were adjusted to meet the input requirements.

2.6 | Plasma ctDNA Analysis

Plasma ctDNA analysis was performed using a hybrid tumor-informed and tumor-agnostic approach that examined both mutation and non-mutation features of ctDNA (K-4CARE, Gene Solutions) [10]. Briefly, cell-free DNA was extracted using the MagMAX Cell-Free DNA Isolation Kit (Thermo Fisher Scientific, USA). For mutation detection, multiplex PCR reactions were used to amplify 5–8 personalized tumor-derived mutations as well as a panel of 700 lung cancer-specific hotspot mutations. Amplified cfDNA fragments were sequenced at an average depth of 100,000X per amplicon. For non-mutation feature detection, cfDNA libraries prepared using the xGen cfDNA Library Prep v2 MC kit (Integrated DNA Technologies, USA) were subjected to shallow whole-genome sequencing (sWGS) at 0.5-1X coverage to determine copy number and fragment length alterations [19]. Tumor fraction (TF) was defined as the percentage of ctDNA in total plasma cfDNA. When mutations were detected, TF was calculated as the mean variant allele frequency (VAF) of all positive mutations. When mutations were absent, TF was defined as the ichorCNA value if positive, or the converted fragment length signal obtained from an in-house linear regression model [19]. To determine limit of detection (LOD), clinical plasma samples with pre-determined tumor fraction (TF) levels of 0.5% were serially diluted to TF levels of 0.1%, 0.05%, 0.025%, and 0.01%, and analyzed in triplicates.

2.7 | Statistical Analysis

Survival curves were estimated by the Kaplan–Meier method, with group differences assessed by the log-rank test. Cox proportional hazard regression was used to calculate the hazard ratio (HR). The Student's t -test was used for comparison of copy number quantification between two panel designs. All statistical analyses were conducted using GraphPad Prism v.10.3.0, with significance set at $p < 0.05$.

3 | Results

3.1 | Assay Design

Current CGP tests using only FFPE DNA profiling had some disadvantages (Figure 1A). For MSI, TMB, and mutations (SNPs, indels), tumor-only analysis is less accurate than tumor-normal analysis; DNA-based fusion detection has lower sensitivity than mRNA-based profiling; and several biomarkers such as CNV, HRD, and germline mutations require separate testing (Figure 1A). Therefore, our assay design integrated FFPE DNA–RNA and LB-ctDNA profiling to address these gaps (Figure 1B). First, DNA from paired FFPE and WBC was hybridized with a 504-gene HDP panel to determine MSI, TMB, mutations (SNPs, indels, common fusions), CNV, and LGR in tumor-normal analysis mode. FFPE DNA libraries were also subjected to shallow WGS to detect CNV at the chromosome level and assess genome instability status for HRD analysis. Second, FFPE

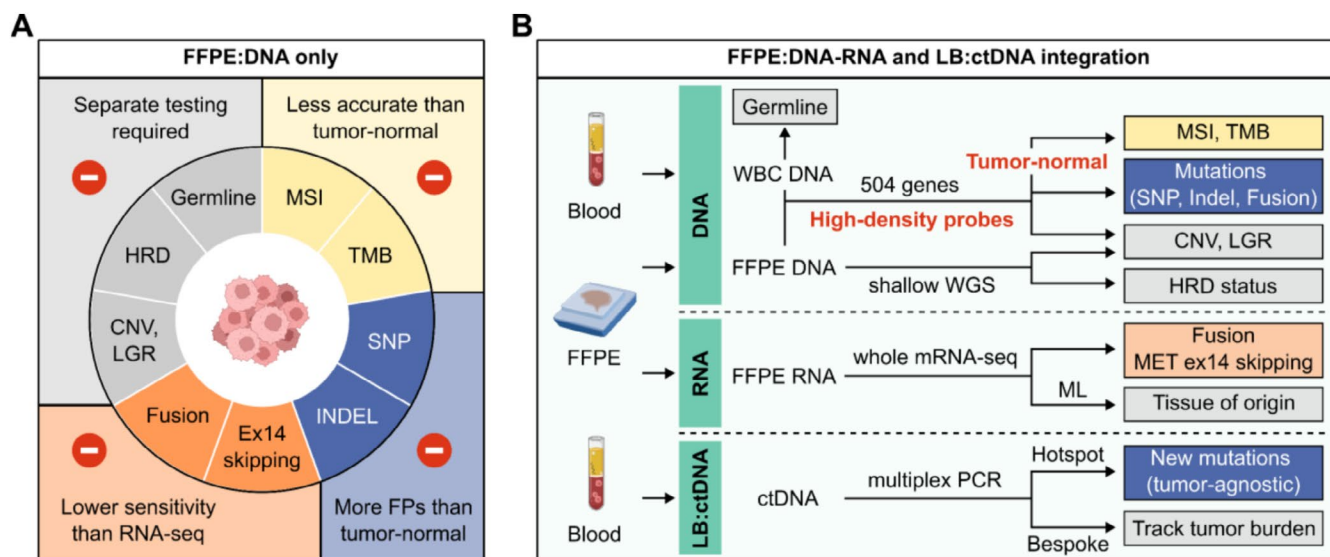


FIGURE 1 | Assay design to integrate FFPE DNA–RNA and LB ctDNA profiling. (A) Disadvantages of CGP using only FFPE DNA to detect biomarkers. (B) Our design to first sequence DNA from matched FFPE and WBC using a 504-gene panel with high-density probes and analyze data in tumor-normal mode to determine microsatellite instability (MSI), tumor mutational burden (TMB), mutations (SNP, Indel, common fusions), copy number variation (CNV), and large genomic rearrangement (LGR). FFPE DNA was also subjected to shallow whole genome sequencing (WGS) to determine CNV, LGR, and homologous recombination deficiency (HRD) status. Second, FFPE mRNA-sequencing was performed to further capture fusion variants, and also combined with machine learning (ML) to predict cancer tissue of origin. Third, plasma ctDNA was subjected to multiplex PCR reactions amplifying both cancer-specific hotspots and personalized tumor-derived mutations to identify new tumor-agnostic mutations and track tumor burden, respectively.

mRNA-sequencing was used to enhance sensitivity for fusion detection and to predict TOO for CUP using an ML model. Third, plasma cfDNA was subjected to hybrid tumor-informed and tumor-agnostic testing. The first multiplex PCR reaction utilized a fixed hotspot panel to identify tumor-agnostic mutations, including resistance mutations. The second reaction used ultra-deep sequencing to track personalized tumor-derived mutations. Together, the status and level of ctDNA helped monitor tumor burden reflecting treatment response (Figure 1B).

3.2 | DNA Sequencing Using High-Density Probes to Detect Copy Number Variation

We compared performance to detect CNV at chromosome-, gene-, and exon- levels between our standard (STD) and high-density probe (HDP) panels. First, SVs—DNA gain/loss at chromosome level, were analyzed in 2 glioma FFPE samples (Figure 2A). The HDP panel showed a clearer signal of chromosome 1p/19q co-deletion in sample TOO289 and detected chromosome 10 loss in sample TOO293 that the STD panel failed to capture. At gene level, signals of gene amplification were also more pronounced using the HDP panel compared to the STD panel (Figure 2B). The deviation of absolute copy number calculated for reference samples was 0.1–0.4 copies for the HDP panel, significantly lower than the variable deviation of up to 1.5 copies for the STD panel (Figure 2B). For gene deletion, we detected both hetero- and homo-zygous co-deletion of *MTAP-CDKN2A* in reference samples using the HDP panel but not the STD panel (Figure 2C). In clinical lung cancer samples ($n = 49$), the prevalence of homozygous loss of *MTAP-CDKN2A* was 12.2% (6/49) (Figure 2C); and 83.3% (5/6) of these samples also had *EGFR* co-mutation (Figure S1A). At exon level, LGRs—insertion/

deletions spanning one or more exons, are challenging to capture with low probe coverage. The STD panel had 81.8% sensitivity to detect LGRs of different sizes in *BRCA1/2* genes, while the HDP panel achieved 100% sensitivity (Figure 2D). In silico simulation of serial tumor fractions showed that LGR detection was robust at VAF of 25% (Figure S1B). Finally, the combination of CNV and *BRCA1/2* mutation detection determined HRD status in clinical samples. In 169 ovarian cancer samples, the prevalence of HRD+ was 55.0% (Figure 2E). Among HRD+ cases, the percentage of those having mutated *BRCA* alone (mBRCA), positive whole genome instability score alone (wGI+), and both mBRCA and wGI(+) together were 11.9%, 58.1%, 30.0% respectively (Figure 2E). *BRCA1* mutations were more prevalent than *BRCA2* mutations (Figure S1C).

3.3 | mRNA Sequencing to Capture Fusion and Predict Tissue of Origin

In fusion-positive reference samples, mRNA profiling achieved 100% sensitivity to capture all fusion variants, while DNA profiling detected 80% of them (Figure 3A). Some variants were not covered by our DNA probe design and hence not detected. Whole mRNA sequencing not only had broader gene coverage but also detected more fusion events than DNA sequencing, explaining its higher sensitivity. However, in real-world clinical samples, we observed a high rate (35%) of FFPE samples having low RNA quality with DV200 below 30% (Figure 3B). Therefore, compared to DNA- or mRNA- profiling alone, the combination of both methods was essential to maximize the sensitivity to detect actionable fusions in clinical samples (Figure 3C). In addition, novel fusion variants reported in the literature could also be identified in clinical samples for research purposes

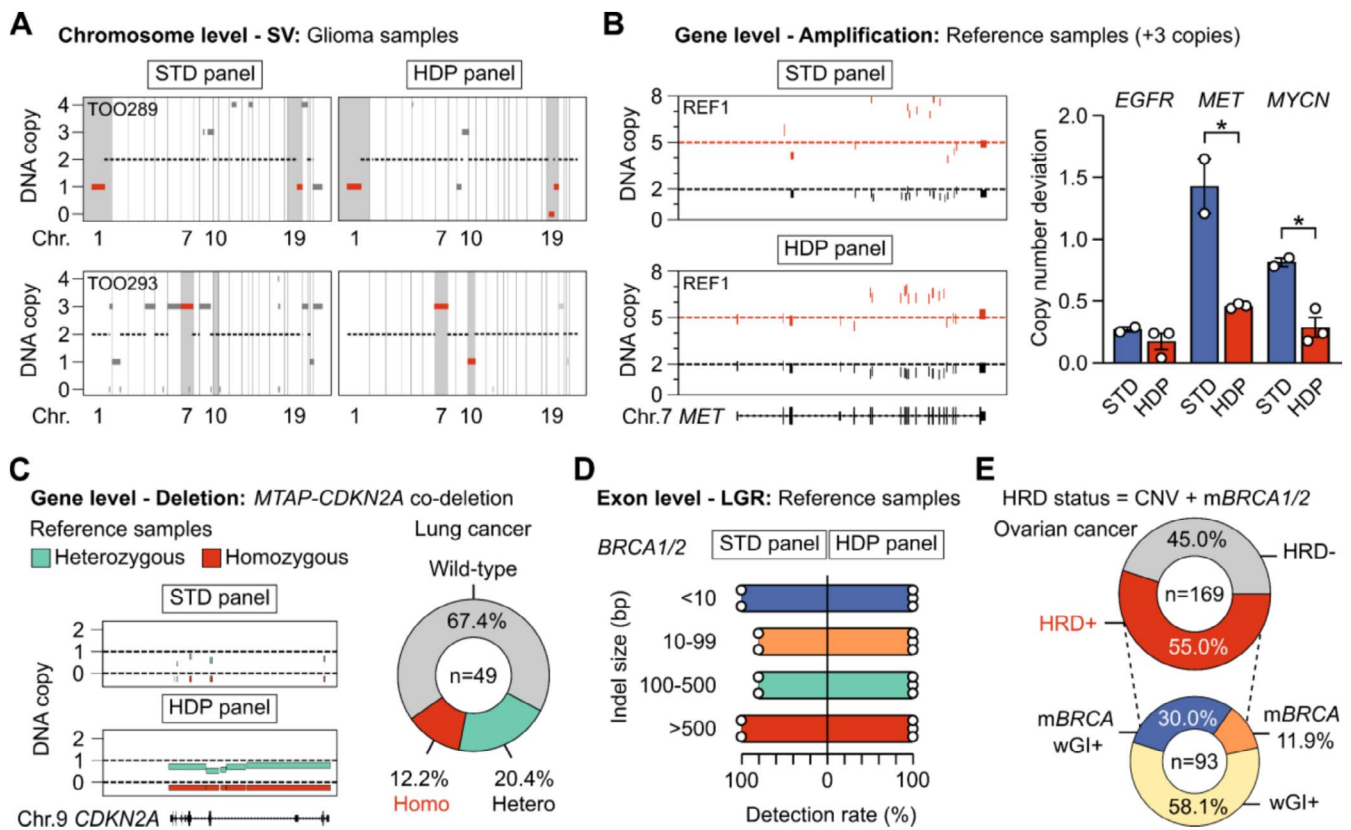


FIGURE 2 | Performance of FFPE DNA sequencing using high-density probes to determine copy number variation. (A) Detection of structural variants (SV), or changes in copy numbers at the chromosome level, in 2 FFPE glioma samples using standard (STD) panel and high-density probe (HDP) panel. (B) Detection of gene amplification and quantification of copy numbers in reference samples using STD and HDP panels ($n=3$). $*p < 0.05$, Student's t -test. (C) Detection of *MTAP-CDKN2A* gene loss in reference samples using STD and HDP panels. The prevalence of homozygous and heterozygous *MTAP-CDKN2A* deletion in lung cancer samples ($n=49$). (D) Detection of large genomic rearrangement (LGR), or large indel of different sizes, in *BRCA1/2* genes in reference samples using STD and HDP panels ($n=20$ LGRs). (E) Determination of homologous recombination deficiency (HRD) status in ovarian cancer samples ($n=169$). HRD+ was defined as either positive whole genome instability (wGI) score, determined by the presence of CNV, and/or mutated *BRCA1/2* (*mBRCA*).

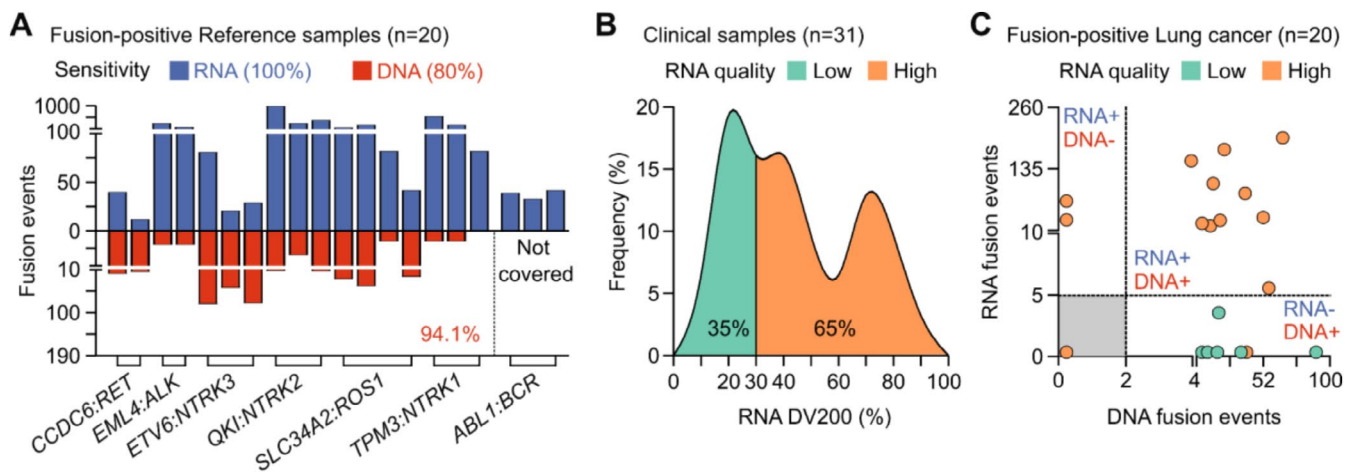


FIGURE 3 | Performance of FFPE DNA and mRNA sequencing to detect fusion variants. (A) Comparison of two methods to detect fusion variants in reference samples ($n=20$ mutations in 10 reference samples). mRNA profiling showed more fusion events and had broader coverage of fusion genes than DNA targeted sequencing. (B) mRNA quality determined by DV200 values in real-world clinical samples ($n=31$). (C) Combining DNA and mRNA sequencing maximized the fusion detection rate for lung cancer FFPE samples ($n=20$).

TABLE 1 | Analytical performance and quality control criteria of the assay.

	Sensitivity ^a	Specificity ^a	Positivity threshold
Analytical performance			
Immunotherapy (FFPE) [10]			
TMB	94%	> 99%	≥ 10 muts/Mb
MSI	R ² correlation with PCR: 97%		≥ 20%
Targeted therapy (FFPE)			
Somatic variants			
DNA-based detection [10]			
SNVs	> 99%	> 99%	≥ 5% VAF ^b
INDELs (< 20 bp)	> 99%	> 99%	
LGR (> 50 bp)	> 99%	> 99%	≥ 25% VAF
Amplification	> 99%	> 99%	≥ 2 copy ratio
Fusion (selected genes)	94%	> 99%	≥ 2 reads at breakpoint
RNA-based detection			
Fusion (all genes)	> 99%	> 99%	≥ 5 reads
<i>MET</i> exon14 skipping	> 99%	> 99%	≥ 10 reads
Hereditary cancer (WBC) [10]			
Germline variants			
SNVs	> 99%	> 99%	N/A
INDELs (< 20 bp)	> 99%	> 99%	N/A
HRD status (FFPE) [12]			
Genomic instability	OPA ^c = 96.3%		wGI ≥ 0
<i>BRCA1/2</i> mutations	> 99%	> 99%	≥ 5% VAF ^b
Quality control criteria			
Interrogated genomic regions			
DNA: coding regions of 504 genes			
RNA: whole coding regions of 19,435 genes			
Tumor content			
≥ 20% (for HRD: ≥ 30%)			
Median depth			
FFPE: 300× and > 90% sites covering at 200×			
WBC: 200× and > 90% sites covering at 100×			
mRNA transcriptome: 30 M total reads			

Abbreviations: FFPE, formalin fixed paraffin-embedded tissue; Indel, insertion/deletion; LGR, large genomic rearrangement; MSI, microsatellite instability; N/A, not available; OPA, overall percent agreement; SNV, single nucleotide variation; TMB, tumor mutational burden; VAF, variant allele frequency; WBC, white blood cells.

^aSensitivity and specificity were determined using standard reference samples: Seraseq gDNA TMB Mix, Seraseq FFPE *BRCA1/2* LGR Reference Material, Seraseq Breast CNV Mix, Seraseq Lung and Brain CNV Mix (SeraCare, USA); Oncospan gDNA, *BRCA* germline I gDNA, Pan-Cancer 6-Fusion Panel, Tru-Q0 DNA Reference Standards (Horizon, USA); EMQN RNA fusion reference samples.

^bSomatic SNVs and INDELs at 1% ≤ VAF < 5% were detected with 82% sensitivity and > 99% specificity.

^cOPA was benchmarked with SOPHiA DDM HRD Solutions (Sophia Genetics, France).

(Figure S2A). Finally, the detection of *MET* ex14 skipping variants also showed a higher sensitivity using mRNA- than DNA-sequencing (Figure S2B). All performance and quality control criteria are listed in Table 1.

For TOO prediction, we first used mRNA transcriptome of TCGA samples as the input for ML model training and selection (Figure 4A, Table S1). Three gene sets: 90 genes from Ye et al. study [15], 2000 genes and 739 genes selected by SIMBA

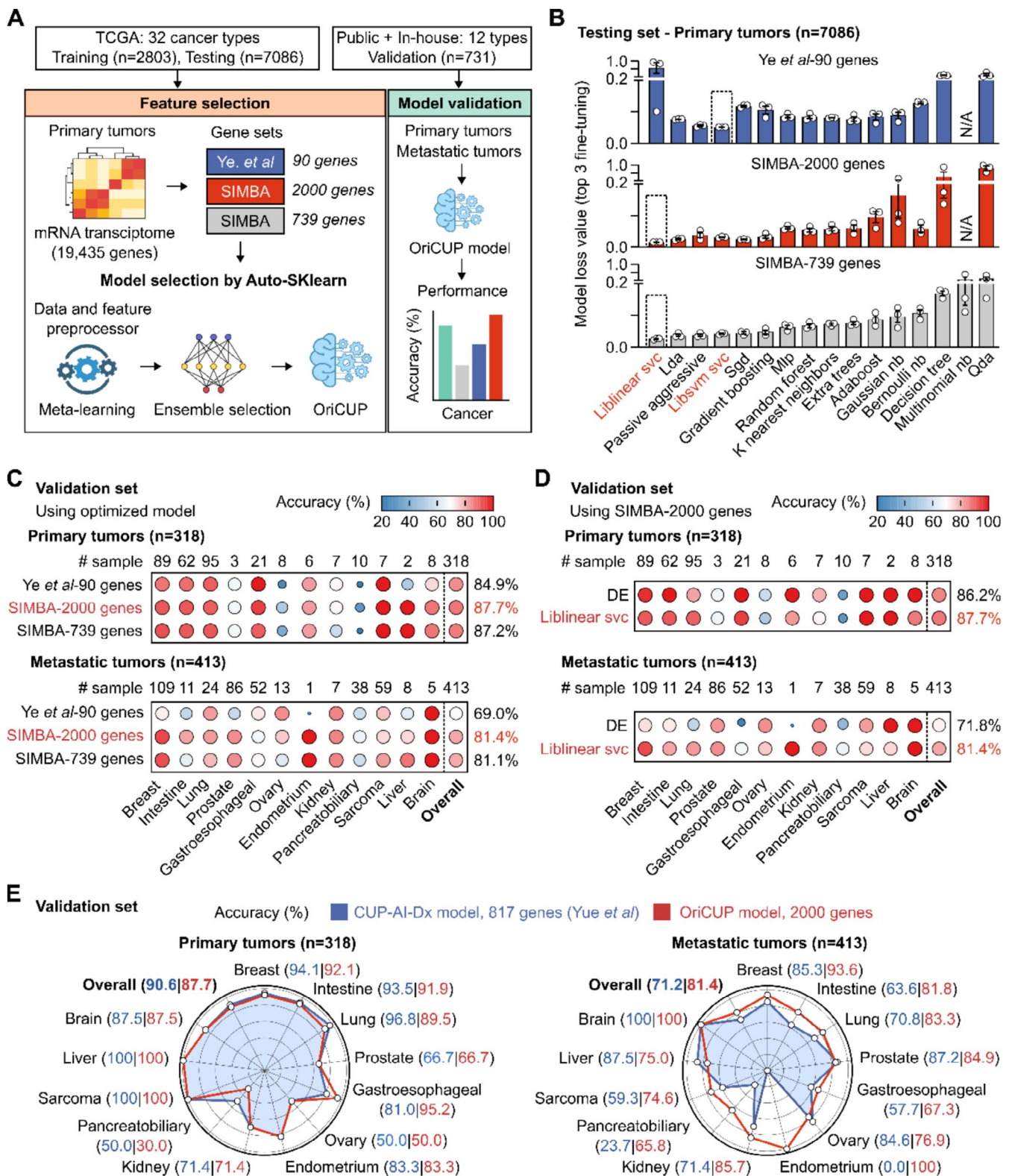


FIGURE 4 | Performance of mRNA sequencing to predict cancer tissue of origin. (A) Workflow to develop OriCUP model to predict tissue of origin (TOO). mRNA profiles of TCGA samples ($n=9889$) and the three gene sets were used as inputs for model training using the Auto-SKlearn framework. Performance was evaluated on an independent validation dataset ($n=731$) consisting of public and in-house data of Gene Solutions (GS). (B) Loss values of different models in the testing dataset after 10-fold cross-validation. The ensemble model having the lowest loss value was selected for each gene set. (C) Performance to predict TOO in the validation dataset using different gene sets and their optimized models. (D) Performance to predict TOO in the validation dataset using the SIMBA-2000 gene set and either its ensemble model or deconvolution (DE) method. (E) The final OriCUP model was benchmarked with the CUP-AI-Dx model in the validation dataset to predict TOO of both primary and metastatic tumors.

method, were used to distinguish 32 types of primary tumors (Figure S3A). Each individual ML pipeline together with pre-processor and feature processor was utilized for hyperparameter optimization via 10-fold cross-validation. The best ensemble model selected based on the lowest loss was Libsvm SVC for 90-gene set, and Liblinear SVC for SIMBA-2000 and SIMBA-739 gene sets (Figure 4B). The sensitivity, specificity and AUC of these models were stable across 10-fold cross-validation (Figure S3B). Among the 3 gene sets, there was no difference in the prediction accuracy using TCGA data with only primary tumors (Figure S3C). In the validation dataset that included both primary and metastatic tumors, the SIMBA-2000 gene set and its optimal Liblinear SVC model showed a slightly higher prediction accuracy than the other 2 sets and hence was selected for the OriCUP model (Figure 4C). Using the same SIMBA-2000 gene set, we also explored deconvolution method, which deconstructed the mixed signals from bulk mRNA sequencing data into individual tissue types based on signature marker genes. Performance of this method was lower than the ML model (Figure 4D). Lastly, we benchmarked the final OriCUP (2000 genes, ensemble model) with the commonly used CUP-AI-Dx (817 genes and 1D-Inception model). Although the performance to predict TOO for primary tumors was not different, the prediction accuracy for metastatic tumors using OriCUP was overall 10% higher than the CUP-AI-Dx (Figure 4E). The difference was most pronounced for lung and gastrointestinal cancers. The overall prediction accuracy of OriCUP was 87.7% for primary tumors and 81.4% for metastatic tumors.

3.4 | ctDNA Monitoring to Identify Tumor-Agnostic Mutations and Predict Disease Progression

We examined the clinical significance of integrating LB-ctDNA with FFPE profiling for a small cohort of 55 stage-IV lung cancer patients. A total of 78 actionable mutations were identified in 52/55 patients, in which FFPE profiling detected 69 mutations

(88.5%) in 50/52 patients, and LB-ctDNA profiling using a hotspot panel specific for lung cancer identified an additional 9 tumor-agnostic mutations (11.5%), including the resistance mutation *EGFR* T790M (Figure 5A). Besides actionable mutations, the total mutation profiles from FFPE were utilized to select the top 4–10 mutations for personalized ctDNA tracking. Among 227 FFPE-derived mutations selected for 55 patients, 114 mutations (50.2%) were detected in the LB of 42 patients (76.4%) (Figure 5B). The extra 14 tumor-agnostic mutations found using the hotspot panel increased the baseline ctDNA detection rate to 87.3% (48 patients) (Figure 5B). In this real-world cohort, 40 patients had serial ctDNA testing before and after treatment initiation with tyrosine kinase inhibitors or immune checkpoint inhibitors. Their clinical response up to 15 months was compared with ctDNA status. Mutation and non-mutation features including CNV and fragmentomics were combined to achieve an LOD of ctDNA detection at 0.01% (Figure S4). Patients who had more than 50% ctDNA decrease from baseline were classified as molecular responders and showed significantly longer progression-free survival (PFS) than patients with <50% ctDNA decrease, or molecular non-responders (HR=9.42, 95% CI: 3.33–26.67, $p < 0.0001$); 12-month PFS were 95.5% and 31.7% respectively (Figure 5C). 84.0% (21/25) of molecular responders achieved clinical complete or partial response, while 93.3% (14/15) of molecular non-responders had confirmed progressive disease.

4 | Discussion

Current CGP testing only examines either DNA or mRNA in isolation and uses either FFPE or LB separately. FFPE DNA remains the cornerstone of CGP assays, but the standard uniform DNA probe design had inadequate sensitivity to detect small CNVs and LGRs [20]. Therefore, increasing probe density along certain chromosomes and small genes, including part of intron regions, has been demonstrated as a promising strategy [21, 22]. In this study, we showed that such HDP design indeed boosted

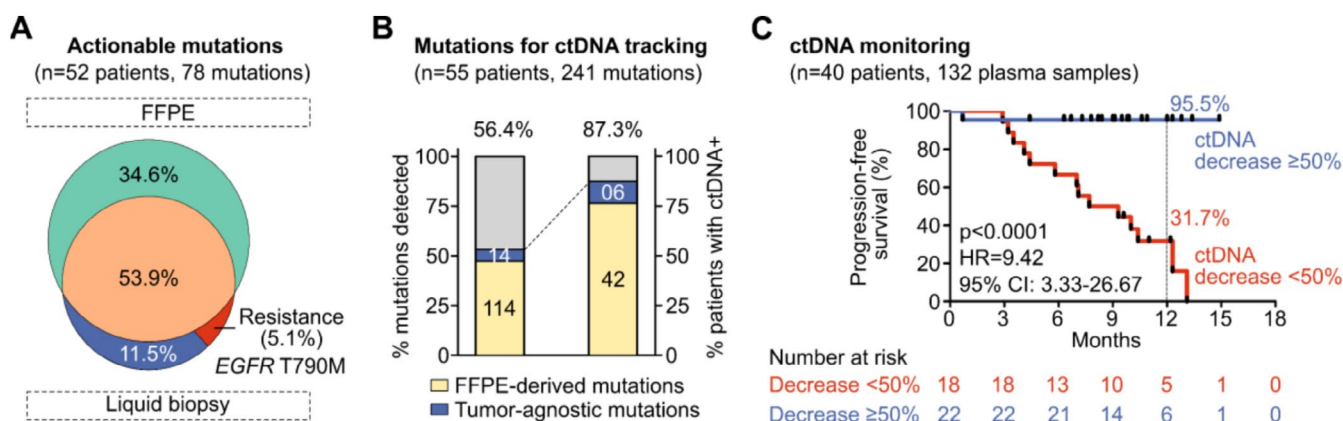


FIGURE 5 | Plasma ctDNA monitoring to identify tumor-agnostic mutations and predict disease progression. (A) In a cohort of 55 stage-IV lung cancer patients, 78 actionable mutations were identified in 52 patients. Venn diagram showing the proportion and overlap of FFPE-derived mutations and liquid biopsy-derived mutations, including resistance mutations. (B) Total mutation profiles (241 mutations) were used for plasma ctDNA monitoring in 55 patients. Bar graphs showing the proportion of FFPE-derived and tumor-agnostic mutations being detected in the plasma, and the percentage of patients having positive ctDNA status before treatment. (C) Kaplan–Meier analysis of progression-free survival (PFS) for 40 patients as stratified by ctDNA dynamics. Patients who had more than 50% ctDNA decrease from baseline showed significantly longer PFS than patients with <50% ctDNA decrease ($p < 0.0001$).

the sensitivity to identify CNVs at both chromosome-, gene-, and exon- levels and improved the accuracy to estimate copy numbers. This is particularly significant for *BRCA1/2* genes, as the prevalence of LGRs in *BRCA1/2* was reported at 2%–3% in breast and ovarian patients [22–25]. Combining CNV and *BRCA1/2* mutation detection, we found that 55.0% of ovarian cancer samples were HRD+, a positive rate similar to prior studies [26, 27]. Furthermore, our HDP panel demonstrated high sensitivity to detect homozygous *MTAP-CDKN2A* co-deletion, an emerging therapeutic target for multiple solid tumors. The prevalence of homozygous *MTAP-CDKN2A* loss in our small lung cancer cohort was 12.2%, comparable with the reported range of 7.6%–13.4% in the literature [28–30], but a larger sample size is required for a more accurate analysis. Besides, our tumor-normal analysis pairing FFPE with WBC DNA helps reduce false-positive findings as it can distinguish germline versus somatic mutations, as well as remove CHIP mutations and sequencing noise [31]. Concurrent germline testing using WBC DNA is also helpful for hereditary cancer risk assessment in high-risk patients [11, 31, 32].

Our results showed that mRNA sequencing was more sensitive to detect fusions as it covered more variants and captured more fusion events than DNA targeted sequencing, consistent with previous reports [3, 33, 34]. This is because mRNA sequencing could detect highly expressed fusion events regardless of their precise genomic breakpoints and even if the underlying DNA rearrangement is complex or cryptic [35]. However, FFPE mRNA is frequently degraded, leading to a high proportion of samples that fail quality control, a common issue reported in developing countries [36]. Therefore, combining DNA- and mRNA-sequencing could leverage the advantages of both methods to maximize patient access to fusion-targeted therapies [37].

The whole mRNA transcriptome could be utilized to predict TOO for CUP, which accounts for ~5% of metastatic cancer cases and often has poor prognosis due to limited treatment options [38]. The overall accuracy of our mRNA-based OriCUP model to predict TOO for metastatic tumors was 81.4%, higher than the mRNA-based CUP-AI-Dx model with 71.2% accuracy [18]. It was also comparable with the deep learning mRNA-based TransCUPtomics model (79.0%) [39], the DNA methylation-based BELIVE model (80.9%) [40] and the popular 90-gene qPCR method (89.2%) [15]. Previous clinical trials using mRNA-based TOO prediction alone to guide site-specific chemotherapy failed to improve survival for patients with CUP [41, 42], but the use of targeted therapies based on actionable mutation profiling alone could modestly prolong PFS in the CUPISCO trial [43]. We speculate that identifying both TOO and actionable mutation profiles by mRNA and DNA profiling could provide stronger guidance and more treatment options for CUP patients. However, the clinical utility of this approach remains to be investigated in future randomized clinical trials.

Although many mutations detected by CGP are not therapeutically actionable, the total mutation profiles could be leveraged as “signatures” of the tumors to detect and quantify plasma ctDNA. With strong evidence from extensive research, the incorporation of LB-ctDNA monitoring has recently been proposed to complement RECIST 1.1 for more precise and earlier prediction of treatment response in metastatic cancer [44]. In addition to

prognostic value, LB-ctDNA profiling could also identify resistance mutations such as *EGFR* T790M to inform therapeutic intervention. However, the required LOD of ctDNA testing and the cutoff of on-treatment ctDNA VAF reduction remain debatable. Our LB-ctDNA test with LOD of 0.01% demonstrated high sensitivity to detect pre-treatment ctDNA and showed a good correlation with clinical response, suggesting its strong potential to be used for treatment monitoring.

Although our integrated approach provides comprehensive and complex genomic data, the high cost of sequencing and analysis can be a significant barrier, limiting its widespread accessibility. The interpretation of complex genomic data requires specialized bioinformatics expertise and clinical knowledge, often leading to challenges in distinguishing between pathogenic alterations and variants of unknown significance, which can complicate clinical decision-making. To streamline the whole process and keep the assay affordable, we prioritized a small number of workflows that concurrently provided multiple biomarkers for multiple cancer types. For instance, sWGS to identify CNV could be coupled with *BRCA* mutation analysis to determine HRD status; targeted sequencing using WBC gDNA allowed better somatic mutation identification using tumor-normal analysis while also providing germline mutation detection. CGP might not be the best option for every patient, but future efforts to reduce costs and increase the number of therapeutically actionable biomarkers would help to narrow the gap between comprehensiveness and affordability.

Although our study has not yet demonstrated clinical utility, which requires future prospective clinical studies, it showed high accuracy and robustness of the unique integrated approach of tumor genomic and transcriptomic profiling, with liquid biopsy ctDNA monitoring. This approach could support better and earlier decision-making to avoid ineffective treatment and unlock more personalized therapeutic options for cancer patients.

Author Contributions

Thien-Phuc Hoang Nguyen, Tien Anh Nguyen, Minh-Duc Nguyen, Ha-Hieu Pham, Minh-Duy Phan, Hoa Giang performed bioinformatics and statistical analysis. Nam H.B. Tran, Van-Anh Nguyen Hoang, Tho Thi Le Vo, My T.T. Ngo processed standard and clinical samples for sequencing. Vinh Quang Bui, Vu Thuong Le, Trong Khoa Mai, Du Quyen Nguyen, Duy Sinh Nguyen, Hoai-Nghia Nguyen performed clinical analysis. Thien-Phuc Hoang Nguyen and Nam H.B. Tran designed experiments and analyzed data. Lan N. Tu conceived the idea, designed experiments, analyzed data, and wrote the manuscript.

Funding

This study was funded by Gene Solutions, Vietnam. The funder did not have any role in the study design, data collection, and analysis, or preparation of the manuscript.

Conflicts of Interest

Nam H.B. Tran, Thien-Phuc Hoang Nguyen, Van-Anh Nguyen Hoang, Tien Anh Nguyen, Minh-Duc Nguyen, Ha-Hieu Pham, Tho Thi Le Vo, My T.T. Ngo, Du Quyen Nguyen, Duy Sinh Nguyen, Hoai-Nghia Nguyen, Minh-Duy Phan, Hoa Giang, Lan N. Tu are current employees of Gene Solutions, Vietnam. The remaining authors declare no conflicts of interest.

Data Availability Statement

The data presented in the study is deposited in the BioProject repository, accession number PRJNA1288126 (<https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA1288126>).

References

1. M. Pankiw, C. Brezden-Masley, and G. S. Charames, "Comprehensive Genomic Profiling for Oncological Advancements by Precision Medicine," *Medical Oncology* 41, no. 1 (2023): 1.
2. W. Mathieson and G. A. Thomas, "Why Formalin-Fixed, Paraffin-Embedded Biospecimens Must be Used in Genomic Medicine: An Evidence-Based Review and Conclusion," *Journal of Histochemistry and Cytochemistry* 68, no. 8 (2020): 543–552.
3. H. Peng, R. Huang, K. Wang, et al., "Development and Validation of an RNA Sequencing Assay for Gene Fusion Detection in Formalin-Fixed, Paraffin-Embedded Tumors," *Journal of Molecular Diagnostics* 23, no. 2 (2021): 223–233.
4. H. Hayashi, Y. Takiguchi, H. Minami, et al., "Site-Specific and Targeted Therapy Based on Molecular Profiling by Next-Generation Sequencing for Cancer of Unknown Primary Site: A Nonrandomized Phase 2 Clinical Trial," *JAMA Oncology* 6, no. 12 (2020): 1931–1938.
5. I. Vanni, L. Pastorino, V. Andreotti, et al., "Combining Germline, Tissue and Liquid Biopsy Analysis by Comprehensive Genomic Profiling to Improve the Yield of Actionable Variants in a Real-World Cancer Cohort," *Journal of Translational Medicine* 22, no. 1 (2024): 462.
6. J. Pascual, G. Attard, F. C. Bidard, et al., "ESMO Recommendations on the Use of Circulating Tumour DNA Assays for Patients With Cancer: A Report From the ESMO Precision Medicine Working Group," *Annals of Oncology* 33, no. 8 (2022): 750–768.
7. M. J. Goldman, B. Craft, M. Hastie, et al., "Visualizing and Interpreting Cancer Genomics Data via the Xena Platform," *Nature Biotechnology* 38, no. 6 (2020): 675–678.
8. D. R. Robinson, Y. M. Wu, R. J. Lonigro, et al., "Integrative Clinical Genomics of Metastatic Cancer," *Nature* 548, no. 7667 (2017): 297–303.
9. V.-A. Nguyen Hoang, N. Nguyen, D. N. le, et al., "Real-World Utilization and Performance of Circulating Tumor DNA Monitoring to Predict Recurrence in Solid Tumors," *JCO Oncology Advances* 2 (2025): e2400084.
10. T. P. Nguyen Hoang, T. A. Nguyen, N. H. B. Tran, et al., "Analytical Validation and Clinical Utilization of K-4CARE: A Comprehensive Genomic Profiling Assay With Personalized MRD Detection," *Frontiers in Molecular Biosciences* 11 (2024): 1334808.
11. V. T. Tran, S. T. Nguyen, X. D. Pham, et al., "Pathogenic Variant Profile of Hereditary Cancer Syndromes in a Vietnamese Cohort," *Frontiers in Oncology* 11 (2021): 789659.
12. T. P. H. Nguyen, N. H. Tran, T. A. Nguyen, et al., "Combination of BRCA Deep Targeted Sequencing and Shallow Whole Genome Sequencing to Detect Homologous Recombination Deficiency in Ovarian Cancer," *medRxiv* (2024): 2024.09.10.24313378.
13. A. Chaubey, S. Shenoy, A. Mathur, et al., "Low-Pass Genome Sequencing: Validation and Diagnostic Utility From 409 Clinical Cases of Low-Pass Genome Sequencing for the Detection of Copy Number Variants to Replace Constitutional Microarray," *Journal of Molecular Diagnostics* 22, no. 6 (2020): 823–840.
14. Y. Liao, G. K. Smyth, and W. Shi, "The R Package Rsubread Is Easier, Faster, Cheaper and Better for Alignment and Quantification of RNA Sequencing Reads," *Nucleic Acids Research* 47, no. 8 (2019): e47.
15. Q. Ye, Q. Wang, P. Qi, et al., "Development and Clinical Validation of a 90-Gene Expression Assay for Identifying Tumor Tissue Origin," *Journal of Molecular Diagnostics* 22, no. 9 (2020): 1139–1150.
16. H. Chen, J. Ryu, M. E. Vinyard, A. Lerer, and L. Pinello, "SIMBA: Single-Cell Embedding Along With Features," *Nature Methods* 21, no. 6 (2024): 1003–1013.
17. Y. Hao, M. Yan, B. R. Heath, Y. L. Lei, and Y. Xie, "Fast and Robust Deconvolution of Tumor Infiltrating Lymphocyte From Expression Profiles Using Least Trimmed Squares," *PLoS Computational Biology* 15, no. 5 (2019): e1006976.
18. Y. Zhao, Z. Pan, S. Namburi, et al., "CUP-AI-Dx: A Tool for Inferring Cancer Tissue of Origin and Molecular Subtype Using RNA Gene-Expression Data and Artificial Intelligence," *eBioMedicine* 61 (2020): 103030.
19. T. Nguyen, V. A. Nguyen Hoang, T. H. Nguyen, et al., "Tumor-Naïve Multimodal Profiling of Circulating Tumor DNA to Detect Minimal Residual Disease in Solid Tumors," *Therapeutic Advances in Medical Oncology* 17 (2025): 17588359251393090.
20. R. Yao, T. Yu, Y. Qing, J. Wang, and Y. Shen, "Evaluation of Copy Number Variant Detection From Panel-Based Next-Generation Sequencing Data," *Molecular Genetics & Genomic Medicine* 7, no. 1 (2019): e00513.
21. Y. Feng, D. Chen, G. L. Wang, V. W. Zhang, and L. J. C. Wong, "Improved Molecular Diagnosis by the Detection of Exonic Deletions With Target Gene Capture and Deep Sequencing," *Genetics in Medicine* 17, no. 2 (2015): 99–107.
22. D. Hua, Q. Tian, X. Wang, et al., "Next-Generation Sequencing Based Detection of BRCA1 and BRCA2 Large Genomic Rearrangements in Chinese Cancer Patients," *Frontiers in Oncology* 12 (2022): 898916.
23. T. Akin Duman and F. N. Ozturk, "Frequency and Distribution of BRCA1/BRCA2 Large Genomic Rearrangements in Turkish Population With Breast Cancer," *Journal of Human Genetics* 68, no. 7 (2023): 485–490.
24. D. H. Kim, H. Chae, I. Jo, et al., "Identification of Large Genomic Rearrangement of BRCA1/2 in High Risk Patients in Korea," *BMC Medical Genetics* 18, no. 1 (2017): 38.
25. W. M. Cao, Y. B. Zheng, Y. Gao, et al., "Comprehensive Mutation Detection of BRCA1/2 Genes Reveals Large Genomic Rearrangements Contribute to Hereditary Breast and Ovarian Cancer in Chinese Women," *BMC Cancer* 19, no. 1 (2019): 551.
26. K. Oda, D. Aoki, H. Tsuda, et al., "Japanese Nationwide Observational Multicenter Study of Tumor BRCA1/2 Variant Testing in Advanced Ovarian Cancer," *Cancer Science* 114, no. 1 (2023): 271–280.
27. C. N. Landen, L. Molinero, H. Hamidi, et al., "Influence of Genomic Landscape on Cancer Immunotherapy for Newly Diagnosed Ovarian Cancer: Biomarker Analyses From the IMagyn050 Randomized Clinical Trial," *Clinical Cancer Research* 29, no. 9 (2023): 1698–1707.
28. P. Ashok Kumar, S. L. Graziano, N. Danziger, et al., "Genomic Landscape of Non-Small-Cell Lung Cancer With Methylthioadenosine Phosphorylase (MTAP) Deletions and Association With EGFR Alterations in NSCLC," *Cancer Medicine* 12, no. 2 (2023): 1157–1166.
29. S. Zhang, W. Liu, M. Wang, et al., "Abstract 1775: The Profiling of MTAP Homozygous Deletion (MTAPhd) in a Chinese Pan-Cancer Cohort," *Cancer Research* 84, no. 6_Supplement (2024): 1775.
30. R. Thummalapalli, D. Muldoon, C. A. Febres-Aldana, et al., "Genomic Landscape of MTAP Deletions and Association With EGFR Alterations in NSCLC," *Journal of Clinical Oncology* 42, no. 16_suppl (2024): 8599.
31. D. Mandelker and O. Ceyhan-Birsoy, "Evolving Significance of Tumor-Normal Sequencing in Cancer Care," *Trends Cancer* 6, no. 1 (2020): 31–39.
32. D. N. Vinh, T. Thi Ngoc Nguyen, T. A. Nguyen Tran, et al., "Breast Cancer Risk Assessment Based on Susceptibility Genes and Polygenic Risk Score in Vietnamese Women," *BJC Reports* 2, no. 1 (2024): 80.

33. C. Heydt, C. B. Wölwer, O. Velazquez Camacho, et al., "Detection of Gene Fusions Using Targeted Next-Generation Sequencing: A Comparative Evaluation," *BMC Medical Genomics* 14, no. 1 (2021): 62.
34. M. A. Socinski, N. A. Pennell, and K. D. Davies, "MET Exon 14 Skipping Mutations in Non-Small-Cell Lung Cancer: An Overview of Biology, Clinical Outcomes, and Testing Considerations," *JCO Precision Oncology* 5 (2021): 663.
35. X. Su, Q. Zheng, X. Xiu, et al., "Challenges and Prospects in Utilizing Technologies for Gene Fusion Analysis in Cancer Diagnostics," *Med-X* 2, no. 1 (2024): 14.
36. H. S. Okuma, H. Yoshida, Y. Kobayashi, et al., "Molecular Pathology Quality Control in Southeast Asia: Results of a Multiregional Quality Assurance Study From MASTER KEY Asia," *Cancer Science* 114, no. 6 (2023): 2664–2673.
37. C. Marchiò, M. Scaltriti, M. Ladanyi, et al., "ESMO Recommendations on the Standard Methods to Detect NTRK Fusions in Daily Practice and Clinical Research," *Annals of Oncology* 30, no. 9 (2019): 1417–1427.
38. L. J. Schipper, K. G. Samsom, P. Snaebjornsson, et al., "Complete Genomic Characterization in Patients With Cancer of Unknown Primary Origin in Routine Diagnostics," *ESMO Open* 7, no. 6 (2022): 100611.
39. J. Vibert, G. Pierron, C. Benoist, et al., "Identification of Tissue of Origin and Guided Therapeutic Applications in Cancers of Unknown Primary Using Deep Learning and RNA Sequencing (TransCUPtomics)," *Journal of Molecular Diagnostics* 23, no. 10 (2021): 1380–1392.
40. S. Zhang, S. He, X. Zhu, et al., "DNA Methylation Profiling to Determine the Primary Sites of Metastatic Cancers Using Formalin-Fixed Paraffin-Embedded Tissues," *Nature Communications* 14, no. 1 (2023): 5686.
41. E. L. Boys, B. Gao, P. Grimison, et al., "Retrospective Analysis of Clinical Characteristics and Outcomes of Patients With Carcinoma of Unknown Primary From Three Tertiary Centers in Australia," *Cancer Medicine* 13, no. 6 (2024): e7052.
42. H. Hayashi, T. Kurata, Y. Takiguchi, et al., "Randomized Phase II Trial Comparing Site-Specific Treatment Based on Gene Expression Profiling With Carboplatin and Paclitaxel for Patients With Cancer of Unknown Primary Site," *Journal of Clinical Oncology* 37, no. 7 (2019): 570–579.
43. A. Krämer, T. Bochtler, C. Pauli, et al., "Molecularly Guided Therapy Versus Chemotherapy After Disease Control in Unfavourable Cancer of Unknown Primary (CUPISCO): An Open-Label, Randomised, Phase 2 Study," *Lancet* 404, no. 10452 (2024): 527–539.
44. M. A. Gouda, F. Janku, A. Wahida, et al., "Liquid Biopsy Response Evaluation Criteria in Solid Tumors (LB-RECIST)," *Annals of Oncology* 35, no. 3 (2024): 267–275.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Table S1:** List of clinical samples and sequencing data used in this study. **Table S2:** List of 504 genes. **Table S3:** List of reference samples used in this study. **Figure S1:** Performance of FFPE DNA sequencing using high-density probes to determine copy number variation. (A) Frequency of homozygous *MTAP-CDKN2A* deletion in lung cancer samples having other actionable mutations ($n=39$). (B) In silico simulation of different VAFs for detection of large genomic rearrangement (LGR) in *BRCA1/2* genes. (C) Percentage of mutated *BRCA1/2* (m*BRCA*) and wild-type *BRCA1/2* (wt*BRCA*) in ovarian cancer samples ($n=169$). **Figure S2:** Performance of FFPE DNA and mRNA sequencing to detect fusion. (A) In fusion-positive clinical samples, mRNA profiling captured more fusion events and had broader coverage of fusion genes and partners than DNA profiling. (B) mRNA profiling could detect MET Ex14 skipping when no DNA mutation was identified in lung cancer samples. **Figure S3:** Performance of mRNA sequencing to

predict cancer tissue of origin. (A) Two-dimensional Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP) using 3 gene sets applied in 32 cancer types in the training dataset revealed distinct clusters corresponding to different cancer types ($n=2803$). (B) Sensitivity and specificity of optimized ensemble models in the testing dataset ($n=6589$) after 10-fold cross-validation across 3 gene sets. The micro-average Receiver Operating Characteristic (ROC) curves from the best cross-validation fold demonstrated stable sensitivity, specificity, and robust discriminative performance in all gene sets. (C) Performance to predict TOO was not different among the 3 optimized ensemble models and corresponding gene sets in the testing dataset ($n=6589$). **Figure S4:** Limit of detection for plasma ctDNA using combined mutation and non-mutation features. Clinical samples were serially diluted to different levels of tumor fractions. Sensitivity to detect mutations and non-mutation genome-wide (GW) features were shown. When both mutation and non-mutation features were combined, limit of detection at 90% confidence (LOD90) was determined at 0.01%.