npj | systems biology and applications

Brief communication

Published in partnership with the Systems Biology Institute

https://doi.org/10.1038/s41540-025-00536-8

Assessment of plasma derived microbiome profiles in lung cancer using targeted and whole exome sequencing

Check for updates

Vichitra Behel[1,2,9], Supriya Hait[2,3,9], Vanita Noronha[1,2], Aniket Chowdhury[2,3], Pratik Chandrani[1,2,4], Vijay Patil[1,2], Nandini Menon[1,2], Rohit Mishra[3], Bhargavi Bawaskar[3], Ganesh Dahimbekar[3], Minit Shah[1,2], Rajiv Kaushal[2,5], Vidya Veldore[6], Hitesh Goswami[6], Atul Bharde[7], Jayant Khandare[7], Gowhar Shafi[7], Anuradha Choughule[1,2], Neetu Tyagi[8], Sanket Desai[2,3], Kumar Prabhash[1,2] ✉ & Amit Dutt[8] ✉

Microbial infections contribute to ~20% of malignancy. Plasma-derived cell-free DNA presents a promising avenue for non-invasive cancer diagnostics, capturing microbial signatures. We analyzed 261 plasma from 50 patients with lung adenocarcinoma by targeted and whole exome sequencing at 10,000 × and 340 × depth, respectively. Comparative analyses of Kraken 2 and IPD2 reveal substantial discrepancies, highlighting challenges in microbial DNA quantification and the need for stringent bioinformatics approaches to ensure accurate cancer microbiome profiling.

Microbial associations with cancer have been a focus of extensive study, with an estimated 20% of cancer cases worldwide linked to infections[1]. The involvement of specific microbes in cancer development is well-established. For instance, *Helicobacter pylori* is implicated in gastric cancer, while *Schistosoma haematobium* is associated with bladder cancer. Similarly, *Fusobacterium nucleatum* has been linked to head and neck cancers as well as colorectal cancer. *Salmonella typhi* and *Human Papillomavirus* (HPV) are associated with gallbladder and cervical cancers, respectively[2–7]. These findings have significantly advanced our understanding of cancer etiology, shedding light on the mechanisms by which microbial infections contribute to cancer initiation and progression while presenting new directions for research and potential therapeutic interventions.

Plasma-derived cell-free DNA (cfDNA) is gaining attention as a promising non-invasive biomarker for disease diagnosis and monitoring. Comprising genetic material from host cells, tumor cells, and microbes, plasma cfDNA offers a unique window into the interaction between cancer and the microbiome[8,9]. Advances in next-generation sequencing (NGS) have enabled the detailed study of microbial cfDNA, uncovering potential microbial signatures linked to cancer.

Despite these advancements, analyzing plasma microbial DNA remains challenging due to its low abundance relative to host DNA. Computational tools like Kraken 2, often used for microbial classification, tend to overestimate microbial burdens due to the non-specific alignment of host reads to microbial databases. Studies, including those by Gihawi et al.[10], highlight the importance of stringent quality control and reliable bioinformatic pipelines to ensure accurate microbial profiling[10,11].

Accurate microbial DNA quantification in plasma could be transformative for cancer diagnostics, especially in cases of unknown primary tumors. This study evaluates plasma microbiome profiling in lung adenocarcinoma using targeted sequencing and whole exome sequencing (WES). By analyzing microbial reads with Kraken 2 and Infectious Pathogen Detector 2 (IPD 2.0), the study compares the strengths and limitations of these methods, providing insights into the feasibility of plasma-based sequencing for cancer microbiome research and tumor identification. Kraken 2 employs k-mer-based classification with probabilistic assignment, while IPD2 integrates host subtraction and reference-guided assembly to minimize misclassification.

To assess the utility of plasma-derived microbiome profiles in lung cancer, we re-evaluated a lung cancer dataset using both targeted and whole exome sequencing. Microbial read counts were estimated using Kraken 2 and compared to those generated by Infectious Pathogen Detector (IPD2), which incorporates host subtraction[5,12]. Our study explored the potential of these sequencing approaches in characterizing the plasma microbiome as an alternative to tissue-based analysis while also evaluating Kraken 2 and IPD 2 for microbial read quantification. We analyzed 261 plasma samples, collected at baseline and during treatment, from 50 patients with histologically

[1]Department of Medical Oncology, Tata Memorial Hospital, Mumbai, Maharashtra, India. [2]Homi Bhabha National Institute, Training School Complex, Anushaktinagar, Mumbai, Maharashtra, India. [3]Integrated Cancer Genomics Laboratory, Advanced Centre for Treatment, Research, and Education in Cancer, Navi Mumbai, Maharashtra, India. [4]Computational Biology, Bioinformatics and Crosstalk Lab, Advanced Centre for Treatment, Research, and Education in Cancer, Navi Mumbai, Maharashtra, India. [5]Department of Pathology, Tata Memorial Hospital, Mumbai, Maharashtra, India. [6]4baseCare Oncosolutions Pvt ltd, Institute of Bioinformatics and Applied Biotech, Bengaluru, Karnataka, India. [7]OneCell Diagnostics, Pune, Maharashtra, India. [8]Integrated Cancer Genomics Laboratory, Department of Genetics, University of Delhi South Campus, New Delhi, India. [9]These authors contributed equally: Vichitra Behel, Supriya Hait. ✉e-mail: kumarprabhashtmh@gmail.com; amitdutt@south.du.ac.in

Table 1 | Comparison of fragment per million counts derived from Infectious Pathogen Detector 2 for all detected microbial genera in the 243 plasma samples analyzed by OncoIndx panel based targeted sequencing with total microbiome burden (TMiB) obtained from Kraken 2

Genus	IPD (read counts)			Kraken (read counts)			Fold change (Kraken/IPD)
	Min	Max	Avg	Min	Max	Avg	
Micrococcus	0.000	0.390	0.107	0.000	1.177	0.244	2.276
Pseudomonas	0.000	0.130	0.025	0.019	3.715	1.042	41.678
Cutibacterium	0.000	0.249	0.047	0.000	0.844	0.161	3.425
Acinetobacter	0.000	0.130	0.024	0.056	2.236	0.589	24.549
Corynebacterium	0.000	0.060	0.006	0.000	0.382	0.138	22.958
Roseomonas	0.000	0.070	0.008	0.000	0.285	0.024	3.001
Aeromonas	0.000	0.300	0.038	0.000	0.200	0.044	1.153
Enterobacter	0.000	0.070	0.005	0.000	1.025	0.128	25.657
Stenotrophomonas	0.000	0.100	0.008	0.048	1.395	0.353	44.066
Achromobacter	0.000	0.050	0.003	0.000	0.233	0.044	14.806
Neisseria	0.000	0.062	0.003	0.000	0.062	0.007	2.176
Staphylococcus	0.000	0.010	0.001	0.000	0.356	0.151	151.021
Shewanella	0.000	0.025	0.001	0.000	0.121	0.029	29.484
Klebsiella	0.000	0.080	0.004	0.000	0.204	0.100	24.986
Paracoccus	0.000	0.062	0.003	0.000	0.520	0.097	32.481
Citrobacter	0.000	0.025	0.001	0.000	0.167	0.028	28.481
Escherichia	0.000	0.030	0.002	0.000	0.318	0.097	48.727
Ralstonia	0.000	0.040	0.002	0.000	0.430	0.101	50.325
Bacillus	0.000	0.020	0.001	0.074	0.876	0.416	416.353

confirmed lung adenocarcinoma at Tata Memorial Hospital, Mumbai. First, we performed targeted sequencing using the OncoIndx gene panel on 243 plasma samples from 32 patients, achieving a median coverage of 10,000× and 72.8 million paired-end reads. These sequencing parameters are comparable to those reported by Zhang et al., who successfully differentiated *H. pylori*- and *EBV*-positive tumors from negative ones by analyzing off-target reads from a panel targeting 425 cancer-related genes[13]. While on-target read analysis indicated a high human genome mapping rate (99.94%), off-target read analysis using IPD2 and Kraken 2 detected microbial signatures, with Kraken 2 reporting significantly higher microbial read counts. Kraken 2 utilized the standard RefSeq database, with a confidence threshold of 0.05, while IPD2 used an in-house curated pathogen database with an FPM (fragments per million) threshold of >1 for viruses and >0 for bacteria. A comparative analysis of microbial read counts generated by IPD2 and Kraken 2 for select genera revealed significant discrepancies, likely influenced by differences in database coverage, algorithmic principles, and classification resolution (Table 1). *Pseudomonas*, for instance, exhibited an average of 0.025 reads using IPD2 compared to 1.042 reads using Kraken 2. *Corynebacterium*, *Achromobacter*, and *Acinetobacter* followed a similar trend, with IPD2 reporting minimal read counts (0.006, 0.003, and 0.024, respectively) in contrast to the substantially higher values obtained from Kraken 2 (0.138, 0.044, and 0.589 reads). The substantially lower microbial read counts from the targeted sequencing dataset can likely be attributed to the sequencing approach and the inherent challenges of detecting low levels of microbial cfDNA in plasma, which may be influenced by physiological barriers like enzymatic degradation, clearance mechanisms, and cell membrane integrity similar to those affecting cfDNA levels[14].

Next, we investigated plasma WES for microbial DNA detection, building on evidence from previous studies that utilized tissue WES[5]. For 18 baseline plasma samples, we achieved a median coverage of 340 × and 125.5 million paired-end reads per sample, with an average of 98.7% of reads mapping to the human genome. Notably, genera with a high total microbiome burden (TMiB)—calculated by normalizing microbial read counts to the total number of sequenced reads analyzed by Kraken 2—included *Pasteurella, Toxoplasma, Nonanavirus, Pseudomonas*, and *Streptomyces*, with prevalence in all samples[5]. Among these, *Pseudomonas* is known to be enriched in lung adenocarcinomas[15]. While *Streptomyces* is a known inhabitant of the human respiratory tract, no associations with lung adenocarcinomas have been reported[16]. Similarly, evidence exists of *Toxoplasma gondii* infection in lung cancer; however, this genus has not been previously reported as part of the lung cancer microbiome[17]. Our Kraken 2-based analysis of plasma WES detected a substantial microbial burden. To validate the differences observed between Kraken 2 and IPD2 methodologies, we performed a direct comparison of bacterial abundance detection across multiple sample types as represented in Supplementary Fig. 1. After normalizing Kraken 2 raw read counts to Fragments Per Million (FPM) for comparable analysis with IPD2 output, we observed that Kraken 2 consistently detected significantly higher bacterial signals across both targeted sequencing ($n = 243$) and whole exome sequencing ($n = 18$) of ctDNA samples. This pattern was evident consistently across the most abundant bacterial genera identified (*Micrococcus*, *Acinetobacter*, and *Pseudomonas*). Wilcoxon matched-pairs signed rank tests confirmed that these differences were statistically significant ($p < 0.0001$ for total bacterial TMiB)

Additionally, we reanalyzed WES data from 51 lung, 27 gallbladder, and 46 head and neck primary tumor biopsies as described in Desai et al.[5] using Kraken 2. We compared Kraken counts from our study and Desai et al.‘s to determine whether different cancers have distinct microbiomes and to compare microbial profiles derived from tumor- and plasma-WES in lung cancer (Fig. 1a). We performed principal component analysis to identify microbial features characteristic of different samples, followed by cluster dendrogram analysis. We observed that distinct microbial features are associated with different types of cancer. Clustering analysis revealed a close resemblance between microbial features identified through plasma-WES and those from tumor-WES in lung cancer patients (Fig. 1b). These similarities suggest that plasma-WES-derived microbial quantification could potentially aid in identifying the primary tumor. This approach could have significant implications for cancers of unknown primary origin, which account for 2–5% of cancers globally and present notable diagnostic and

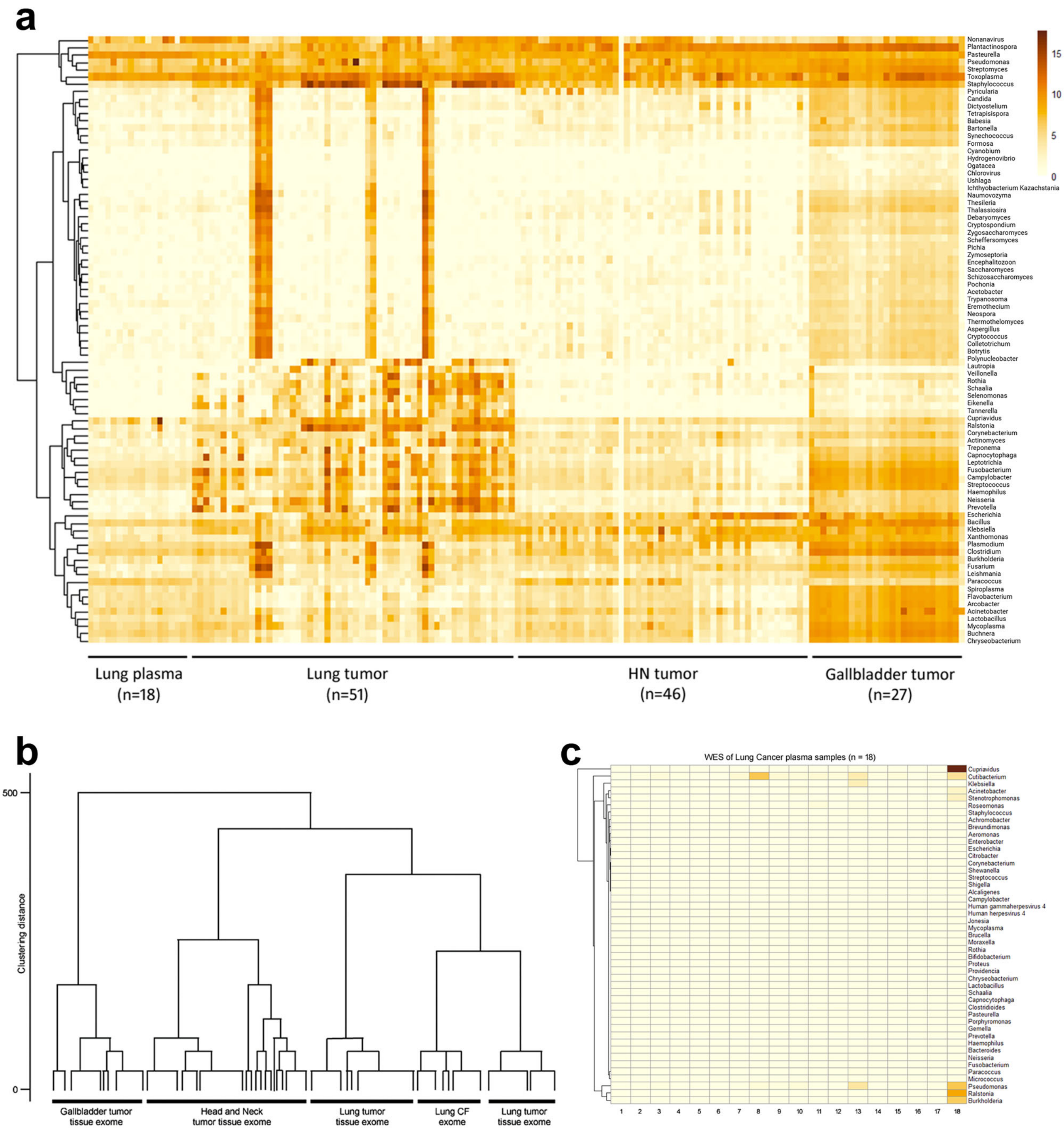

Fig. 1 | Microbiome analysis in cancer samples using whole exome sequencing. **a** Kraken-derived total microbiome burden of commonly detected microbial genera in plasma samples from lung cancer patients and tumor samples from lung, head, and neck, and gall bladder cancer patients. **b** Cluster dendrogram based on principal component analysis of genus-level microbial quantification data from Kraken, stratifying 51 lung tumors, 27 gall bladder tumors, 46 head and neck tumors, and 18 plasma samples from lung cancer patients. **c** Microbial profiles of lung cancer patients obtained from plasma whole exome sequencing using the Infectious Pathogen Detector 2.

therapeutic challenges, as previously reported[18,19]. However, when we re-analyzed the microbial profiles from plasma WES using IPD2, we observed that while IPD2 detected microbes in 17 out of 18 samples, the microbial burden was relatively lower than that identified using Kraken 2. The FPM counts—representing the number of paired-end fragments aligned to pathogens, divided by the total paired reads—ranged from 0.0048 to 4.85 for microbes detected in plasma (Fig. 1c). These counts undergo additional normalization, transforming fragment counts into FPM for each pathogen. This process involves calculating fragments in a feature multiplied by 10^6, divided by the total number of fragments sequenced for a sample, and further dividing FPM by the pathogen's genome length to yield fragments per kilobase of transcript per million mapped reads. The normalization method utilizes the total number of reads sequenced in a sample as the denominator, with the counts of fragments assigned to a pathogen genome (or genus) being adjusted to a per-million scale, ensuring comparability across samples. For all genera detected by IPD2, we compared the FPM counts with Kraken-derived TMiB (Table 2). The data suggest that Kraken 2-based TMiB can be exaggerated by up to 15 times compared to IPD2, necessitating stringent bioinformatics pipelines to minimize false positives. This indicates that IPD2 may help reduce the occurrence of false positives.

Table 2 | Comparison of fragment per million counts derived from Infectious Pathogen Detector 2 for all detected microbial genera in 18 plasma samples analyzed by whole exome sequencing with total microbiome burden obtained from Kraken 2

Genus	IPD (read counts)			Kraken (read counts)			Fold change (Kraken/IPD)
	Min	Max	Avg	Min	Max	Avg	
Cupriavidus	0.000	4.855	0.272	0.008	374.157	20.97	77.07
Cutibacterium	0.000	0.890	0.177	0.000	0.17488	0.039	0.21
Pseudomonas	0.000	0.923	0.118	0.069	4.536	0.561	4.737
Ralstonia	0.000	1.351	0.077	0.000	6.123	0.362	4.681
Burkholderia	0.000	0.870	0.048	0.021	0.601	0.082	1.700
Acinetobacter	0.000	0.245	0.034	0.060	0.718	0.168	4.968
Roseomonas	0.000	0.119	0.027	0.000	0.075	0.020	0.731
Klebsiella	0.000	0.363	0.025	0.009	0.081	0.032	1.275
Stenotrophomonas	0.000	0.289	0.020	0.004	0.351	0.074	3.704
Staphylococcus	0.000	0.089	0.016	0.038	0.142	0.075	4.523
Paracoccus	0.000	0.074	0.013	0.145	0.409	0.287	22.049
Micrococcus	0.000	0.059	0.011	0.004	0.321	0.072	6.468
Brevundimonas	0.000	0.030	0.005	0.092	2.160	0.391	81.326
Aeromonas	0.000	0.021	0.005	0.000	0.083	0.020	4.421
Achromobacter	0.000	0.042	0.004	0.000	0.238	0.031	7.133
Escherichia	0.000	0.024	0.003	0.004	0.056	0.022	8.217
Enterobacter	0.000	0.034	0.002	0.000	0.121	0.020	10.500
Citrobacter	0.000	0.024	0.001	0.000	0.021	0.009	6.658
Corynebacterium	0.000	0.021	0.001	0.000	0.146	0.035	30.037
Shewanella	0.000	0.021	0.001	0.008	0.058	0.024	20.549
Streptococcus	0.000	0.019	0.001	0.015	0.075	0.044	42.434
Shigella	0.000	0.005	0.000	0.000	0.008	0.002	7.759

The principal component analysis demonstrated clear clustering of samples based on the classification method used, with Kraken 2 derived profiles showing greater within-group variability compared to the more consistent IPD2 profiles. These differences highlight the significant impact that bioinformatic tool selection can have on microbial detection in plasma samples. While WES is not specifically designed to detect microbial DNA, it may occasionally identify microbial sequences due to nonspecific binding or sample complexity. However, the biological significance of these detected microbes in the context of cancer remains uncertain.

The considerable differences observed between Kraken 2 and IPD2 results can be attributed to three major factors. First, the database composition critically impacts classification outcomes, with Kraken 2's extensive reference database increasing sensitivity but potentially introducing false positives from environmental contaminants and incomplete genome assemblies. In contrast, IPD2's curated pathogen database employs more stringent filtering, resulting in fewer but potentially more reliable microbial identifications. Second, the algorithmic principles differ fundamentally: Kraken 2's k-mer-based probabilistic classification approach can lead to inflated microbial read counts through non-specific alignments, while IPD2 integrates host subtraction and reference-guided assembly to minimize false positives. Third, the classification resolution varies, with Kraken 2 offering broader taxonomic coverage but potentially overestimating microbial burden, whereas IPD2 applies stricter criteria that reduce noise in microbial detection. These findings underscore the importance of selecting appropriate bioinformatic tools for low-biomass samples like plasma, where stringent filtering approaches may be crucial for accurate microbial profiling in cancer research.

This study has several limitations, including a limited sample size, the absence of negative controls, and the potential for contamination during sample processing. Future studies with larger cohorts and stringent contamination controls are necessary to validate these findings. Our findings highlight the critical importance of stringent quality control measures when using sequencing approaches aimed at capturing host DNA for microbial quantification. Reference-guided assembly coupled with the filtration of low-quality reads is essential to mitigate errors arising from database contamination, a known issue that can significantly inflate microbial burden estimates. Despite employing highly accurate tools like Kraken, misclassification of human reads as microbial due to contaminated microbial genome databases remains a substantial challenge[10]. Also, we did not perform repeated detection analyses on samples in assessing method reproducibility. Future studies should incorporate technical replication to establish reliability and confidence intervals for microbial detection in plasma cfDNA.

Our analyses suggest that targeted sequencing and whole exome sequencing may not be optimal methods for assessing cancer-associated microbial burden in plasma. We emphasize the need for robust quality control to minimize inflated microbial counts when utilizing host subtraction methods. Additional discussion regarding our microbial analysis can be found in the Supplementary Note.

Methods

General study details

This is a single-center, prospective study conducted among adult patients with histologically confirmed lung adenocarcinoma receiving treatment at the Tata Memorial Hospital, Mumbai, India. Fifty patients were enrolled. The study was approved by the Institutional Ethics Committee of the Tata Memorial Hospital, and written informed consent was obtained from all eligible participants. It was conducted in accordance with the Declaration of Helsinki and the ethical standards of the Institutional Ethics Committee on human research.

Patient sample collection and processing

A total of 261 plasma samples were collected from 50 patients with lung adenocarcinoma at baseline and several other time points along the course of treatment. The studies involving humans were approved by The Institutional Ethics Committee (IEC) of Tata Memorial Centre (IEC study number: 900233). Written informed consent was obtained from all participants, and the studies were conducted in accordance with the local legislation and institutional requirements. Targeted sequencing was performed for a total of 243 plasma samples from 32 patients who were followed longitudinally throughout their treatment. For the remaining 18 patients, WES was performed for baseline plasma samples. Plasma was separated from 10 mL of peripheral blood and collected in dipotassium ethylenediaminetetraacetic acid vacutainer tubes within 6 h of blood draw by centrifugation at 2000 × *g* for 20 min followed by 3200 × *g* for 30 min at room temperature. Cell free DNA (cfDNA) was isolated from plasma using the QIAamp MinElute ccfDNA Mini Kit (Catalog no. 55204, Qiagen, Hilden, Germany) per the manufacturer's instructions. The isolated cfDNA was quantified using the Qubit 4 Fluorometer (Invitrogen™, Waltham, USA), and its quality was assessed using the Agilent 4200 TapeStation (Agilent, USA).

Targeted sequencing of plasma samples

Libraries were prepared using 10–50 ng of input cfDNA. DNA fragment ends were repaired, followed by A-tailing, to produce 5'-phosphorylated and 3'-dA-tailed fragments and adapter ligation. Following this, the libraries were purified using AMPure XP beads and amplified by polymerase chain reaction (PCR). The SureSelect XT HS2 DNA System (Agilent Technologies, Santa Clara, CA, USA) was used with a custom-designed panel, OncoIndx to profile cancer-relevant genes, was used for target enrichment. This was followed by hybrid capture using streptavidin-coated magnetic beads. The captured libraries were amplified by PCR and purified using AMPure XP beads. The quantity and quality of the library DNA were assessed using Qubit 4 Fluorometer (Invitrogen™, Waltham, USA) and Agilent 4200 TapeStation (Agilent, USA), respectively. Qualified libraries were sequenced using the NextSeq™ Systems (Illumina Inc., CA, USA) to generate 150-bp paired-end reads at a median coverage of 10,000 ×.

WES of plasma samples

Libraries for WES were prepared using at least 25 ng of cfDNA. Following end repair, A-tailing, and adapter ligation, the libraries were purified using AMPure XP beads and amplified by PCR. The SSELXT HS Human All Exon V8 kit (Catalog no. 5191-6874, Agilent Technologies, CA, USA) was used for target enrichment, followed by hybrid capture. The captured libraries were amplified, purified, and sequenced using the NextSeq™ Systems (Illumina Inc., CA, USA) to generate 150-bp paired-end reads at a median coverage of 340×.

Contamination control measures

To minimize and account for contamination, strict protocols were implemented throughout the study. All sample processing was conducted in a dedicated clean room environment with HEPA filtration. Before any procedure, all work surfaces were decontaminated using 70% ethanol. Only sterile, DNA/RNA-free consumables and reagents were used for sample processing. At the data analysis level, two-tier filtration approach using IPD 2 was specifically implemented to distinguish true microbial signals from background contamination.

Microbial quantification from targeted sequencing and WES data

FastQC was used to assess the quality of FASTQ files generated by NGS[20]. Leading and/or trailing bases with low quality were trimmed using the head crop method, and reads with a significant proportion of N bases were removed using Trimmomatic. Duplicate reads were removed, and the remaining reads were mapped to the human genome reference sequence GRCh38 (gencode v30) using the Burrows-Wheeler Aligner v2[21]. Confirmation of known oncogenic driver mutations such as the epidermal growth factor receptor (EGFR) L858R and exon19 deletions was performed using the Integrative Genome Viewer (IGV v2.8.2)[22]. Unmapped reads were mapped against the microbial reference genome to obtain microbial quantification data. For microbial classification, two different classification tools were employed, Kraken 2 (version 2.1.1) was used with a standard microbial database to quantify reads from bacteria, viruses, fungi, and archaea in plasma samples from RefSeq[11]. The classification was performed with default parameters, threshold was set to a minimum k-mer count of 1. Similarly, IPD 2, we used version 2.0 with the comprehensive pathogen database that integrates sequences from NCBI RefSeq, GenBank, and specialized pathogen databases of known pathogenic bacteria and viruses used for the detection of known pathogenic bacteria and viruses in plasma samples[12]. IPD 2 provides a normalized count of reads that align to the microbial reference genome as fragments per million (FPM). An FPM count of >1 was considered to be a positive signal for the detection of viruses, while that of >0 was considered to be a positive signal for detection of bacteria.

Additionally, we reanalyzed the WES data of 51 lung, 27 gall bladder, and 46 head and neck tumors from Desai et al.'s study using Kraken[5]. Kraken microbial counts for plasma samples from our study and tissue samples from Desai et al.'s study were used to assess whether different cancer types have distinct microbial profiles. DESeq2-based raw read count normalization was performed on the dataset using the variance stabilization technique (VST), followed by principal component analysis (PCA) using the prcomp function of the R package to identify principal features from the count matrix data. Principal features identified by PCA were subject to hierarchical clustering using the FactoMineR package[23].

Data availability

Data availability ArrayExpress database accession numbers for the datasets of tumor samples: E-MTAB-11412, E-MTAB-9766, E-MTAB-6619, E-MTAB-4653, E-MTAB-8801, E-MTAB-11404, E-MTAB-9281 and E-MTAB-11407.The liquid biopsy datasets generated during and/or analyzed during the current study are not publicly available due to an ongoing analysis related to this study but are available from the corresponding author on reasonable request with adequate justification.

Code availability

The analysis utilized publicly available Kraken 2 and in-house IPD2 tools (https://github.com/sanket-desai/InfectiousPathogenDetector2).

Received: 15 February 2025; Accepted: 13 May 2025;
Published online: 07 June 2025

References

1. Rositch, A. F. Global burden of cancer attributable to infections: the critical role of implementation science. *Lancet Glob. Health* **8**, e153–e154 (2020).
2. Jacqueline, C. et al. Infections and cancer: the "fifty shades of immunity" hypothesis. *BMC Cancer* **17**, 257 (2017).
3. Aziz, R. K., Khalifa, M. M. & Sharaf, R. R. Contaminated water as a source of Helicobacter pylori infection: a review. *J. Adv. Res.* **6**, 539–547 (2015).
4. Mostafa, M. H., Sheweita, S. A. & O'Connor, P. J. Relationship between schistosomiasis and bladder cancer. *Clin. Microbiol Rev.* **12**, 97–111 (1999).
5. Desai, S. et al. Fusobacterium nucleatum is associated with inflammation and poor survival in early-stage HPV-negative tongue cancer. *NAR Cancer* **4**, zcac006 (2022).
6. Castellarin, M. et al. Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Res.* **22**, 299–306 (2012).
7. Iyer, P. et al. Non-typhoidal Salmonella DNA traces in gallbladder cancer. *Infect. Agent Cancer* **11**, 12 (2016).
8. Glyn, T. & Purcell, R. Circulating bacterial DNA: a new paradigm for cancer diagnostics. *Front. Med.* **9**, 831096 (2022).
9. Kowarsky, M. et al. Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA. *Proc. Natl. Acad. Sci. USA* **114**, 9623–9628 (2017).
10. Gihawi, A. et al. Major data analysis errors invalidate cancer microbiome findings. *mBio* **14**, e0160723 (2023).

11. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
12. Desai, S. et al. An integrated approach to determine the abundance, mutation rate and phylogeny of the SARS-CoV-2 genome. *Brief. Bioinform.* **22**, 1065–1075 (2021).
13. Zhang, X. et al. Distinct genomic profile in h. pylori-associated gastric cancer. *Cancer Med.* **10**, 2461–2469 (2021).
14. Khier, S. & Lohan, L. Kinetics of circulating cell-free DNA for biomedical applications: critical appraisal of the literature. *Fut. Sci. OA* **4**, FSO295 (2018).
15. Greathouse, K. L. et al. Interaction between the microbiome and TP53 in human lung cancer. *Genome Biol.* **19**, 123 (2018).
16. Lara, A. C. et al. The genome analysis of the human lung-associated Streptomyces sp. TR1341 revealed the presence of beneficial genes for opportunistic colonization of human tissues. *Microorganisms* **9**, https://doi.org/10.3390/microorganisms9081547 (2021).
17. Bajnok, J. et al. High frequency of infection of lung cancer patients with the parasite Toxoplasma gondii. *ERJ Open Res.* **5**, https://doi.org/10.1183/23120541.00143-2018 (2019).
18. Lee, M. S. & Sanoff, H. K. Cancer of unknown primary. *BMJ* **371**, m4050 (2020).
19. Hillen, H. F. Unknown primary tumours. *Postgrad. Med. J.* **76**, 690–693 (2000).
20. Andrews, S. A quality control tool for high throughput sequence data. [Software]. Retrieved from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).
21. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
22. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
23. Lê, S., Josse, J. & Husson, F. FactoMineR: An R Package for Multivariate Analysis. *J. Stat. Softw.* **25**, 1–18 (2008).

Acknowledgements

The authors would like to thank all the technical staff of the Department of Medical Oncology and Department of Pathology, Tata Memorial Hospital/ ACTREC, Mumbai, India, who assisted in this study. This research was supported by the Indian Council of Medical Research (ICMR), Government of India, under Grant 01/ICMRCAREP-2023-0000278/AD/DevDiv. We gratefully acknowledge ICMR for their financial support, which was instrumental in conducting this study.

Author contributions

Study conception and design: V.B., S.H., S.D., K.P., A.D. Data collection: Sample collection: V.B., P.C., V.P., N.M., M.S., R.K., A.C., V.N., K.P.; Whole exome sequencing: V.V., H.G.; ctDNA sequencing: A.B., J.K., G.S. Analysis and interpretation of results: V.B., S.H., A.C., R.M., B.B., G.D., N.T., S.D., A.D. Draft manuscript preparation: V.B., S.H., S.D., A.D. All authors reviewed the results and approved the final version of the manuscript: V.B., S.H., V.N., A.C., P.C., V.P., N.M., R.M., B.B., G.D., M.S., R.K., V.V., H.G., A.B., J.K., G.S., A.C., N.T., S.D., K.P., A.D.

Funding

Open access funding provided by Department of Atomic Energy.

Competing interests

V.V. and H.G. is employed by 4baseCare Oncosolutions Pvt Ltd. A.B., J.K. and G.S. are employed by OneCell Diagnostics Pvt Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41540-025-00536-8.

Correspondence and requests for materials should be addressed to Kumar Prabhash or Amit Dutt.