Cancer Type Classification Using Plasma Cell Free RNAs Derived From Human and Microbes

Shanwen Chen

Peking University First Hospital

Yunfan Jin

Tsinghua University

Siqi Wang

Tsinghua University

Shaozhen Xing

Tsinghua University

Yingchao Wu

Peking University First Hospital

Yuhuan Tao

Tsinghua University

Yongchen Ma

Peking University First Hospital

Shuai Zuo

Peking University First Hospital

Xiaofan Liu

Tsinghua University

Yichen Hu

Nanchang University

Hongyan Chen

Chinese Academy of Medical Sciences and Peking Union Medical College

Yuandeng Luo

The First Hospital Affiliated to Army Medical University

Feng Xia

The First Hospital Affiliated to Army Medical University

Chuanming Xie

The First Hospital Affiliated to Army Medical University

Jianhua Yin

Navy Medical University

Xin Wang

Chinese Academy of Medical Sciences and Peking Union Medical College

Zhihua Liu

Chinese Academy of Medical Sciences and Peking Union Medical College

Ning Zhang

Peking University First Hospital

Zhenjiang Zech Xu

Nanchang University

Zhi John Lu

Tsinghua University

Pengyuan Wang (✉ pengyuan_wang@bjmu.edu.cn)

Peking University First Hospital <https://orcid.org/0000-0002-1210-4056>

Research

Keywords: Liquid biopsy, Microbiome, Cancer classification, Biomarker, Cell-free RNA

Posted Date: October 29th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1012781/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: The utilities of cell free nucleic acids in monitoring cancer have been recognized by both scientists and clinicians. In addition to human transcripts, a fraction of cell free nucleic acids in human plasma were proved to derived from microbes, and reported to have some relevance to cancer.

Methods: To get a better understanding of plasma cell free RNAs (cfRNAs) in cancer patients, we profiled cfRNAs in ~300 plasma samples of five cancer types (colorectal cancer, stomach cancer, liver cancer, lung cancer, esophageal cancer) and healthy donors with RNA-seq.

Results: Microbe derived cfRNAs were consistently detected by different computational methods when potential contaminations were carefully filtered. Clinically relevant signals can be identified from human and microbial reads, and alteration in human cfRNA expression and virus abundance both suggests some cancer patients were immunosuppressed, as indicated by enriched KEGG pathways of downregulated human genes and higher prevalence torque teno virus. Our data supports the diagnostic value of human and microbe derived plasma cfRNAs for cancer detection, as an area under receiver operating characteristic (ROC) curve of 0.931 for distinguishing cancer patients from healthy donors was achieved on validation set, using both human and microbial features. Moreover, these cfRNAs both have some cancer type specificity, and could distinguish tumors of different primary locations. Compared to using human feature alone, combining human and microbial features improves the average validation accuracy of between cancer type classification by 11.5%.

Conclusions: In summary, this work provides evidence for the clinical relevance of human and microbe derived plasma cfRNAs, and their potential utilities in cancer detection, and determination of tumor sites.

Introduction

Lately non-invasive liquid biopsy of plasma cell free nucleic acids is emerging as a convenient and cost-effective method for cancer screening and monitoring. The clinical utilities of cell free DNA (cfDNA) and cell free RNA (cfRNA) in cancer have been extensively studied. Mutations^[1], methylation level^[2], fragmentation patterns^[3] of plasma cfDNA, and expression level of different cfRNA species (miRNA, circular RNA, srpRNA, lncRNA, mRNA, etc)^[4-6] in plasma, platelets, and extracellular vesicles (EVs) were identified as potential diagnostic or prognostic markers. In addition to early detection, it's also favorable if liquid biopsy could provide clues about tumor's primary location, to guide further clinical decisions. Plasma cfDNA methylation and platelet transcriptome were reported to have some cancer type specificity^[2, 4], but whether plasma cfRNAs have such property remains largely uncharacterized.

The studies of human cancer related microbiome are increasingly valued for their novel biological insights and potential clinical applications. It is well established that some bacteria and viruses are involved in cancer development and progression. For instance, chronic infection of hepatitis B virus (HBV) and human papillomavirus (HPV) is the leading cause of liver cancer and cervical cancer, respectively^[7, 8].

Helicobacter pylori infection is a well-known risk factor for developing gastric cancer^[9]. *Fusobacterium nucleatum* was reported to drive tumorigenesis in colon cancer^[10]. It's also reported that in pancreatic cancer, higher microbial diversity predicts better prognosis^[11]. A more recent study reported that cancer type-specific living bacteria can be detected inside tumor cells, suggesting there are unexpectedly complicated interactions between microbes and tumor cells^[11].

Traditionally, blood was believed to be sterile in individuals without sepsis^[12, 13]. Although it remains controversial whether blood of healthy donor contains living bacteria^[4, 14], several recent studies suggested that bacteria derived nucleic acids could be confidently detected in human plasma, which cannot be simply attributed to contaminations in reagents and other potential sources^[12, 15–17]. Many uncharacterized bacteria and viruses could be assembled from blood DNA-seq data^[16]. It was also reported that in obese patients, gut microbes derived EVs, which contain microbial DNA, could enter bloodstream, and induce inflammation response^[18]. A recent study also suggested that the abundance of microbial derived plasma cfDNA could accurately distinguish between different cancer types^[19].

Total RNA sequencing captures RNA fragments regardless of their originations. To study the biological relevance and cancer type specificity of human and microbe derived cfRNAs, we investigated diverse cfRNA species (>50nt, rRNA depleted) in about 300 plasma samples of cancer patients and healthy donors (HDs). This cohort includes five cancer types (colorectal cancer, stomach cancer, liver cancer, lung cancer and esophageal cancer) that were responsible for 75% of cancer-related mortality in China^[20]. Most of the cancer patients were in early stages. As far as we know, our study demonstrated for the first time that both human and microbe derived RNAs in plasma detected by cfRNA-seq could reflect cancer type specific information. We also showed that combining microbial cfRNA signatures could improve the performance of human cfRNAs in cancer detection and classification.

Materials And Methods

Cohort design and sample collection

The cohort in this study included 295 plasma samples in total. Except for 65 previously published samples (GSE142987: 35 liver cancer patients and 30 healthy donors)^[21], we sequenced the total cfRNAs (>50nt) in 230 additional plasma samples (54 colorectal cancer, 37 stomach cancer, 27 liver cancer, 35 lung cancer, 31 esophageal cancer and 46 HDs).

Samples were obtained between October 2018 to January 2020 from 6 clinical centers in China: Peking University First Hospital (PKU, Beijing), Peking Union Medical College Hospital (PUMCH, Beijing), Department of Epidemiology Navy Medical University (ShH-1, Shanghai), Eastern Hepatobiliary Surgery Hospital (ShH-2, Shanghai), National Center for Liver Cancer (ShH-3, Shanghai) and Southwest Hospital (SWU, Chongqing). The study was approved by the local institutional research ethics committees. Informed consent was obtained from all patients.

Peripheral whole blood samples were collected from participants before therapies using EDTA-coated vacutainer tubes. The tubes were inverted 8-10 times to mix blood with anticoagulant. Plasma was separated by a standard clinical blood centrifugation protocol within 2 hours after collection. All plasma samples were aliquoted and stored at -80°C before cfRNA extraction.

cfRNA-seq library preparation

Cell free RNAs (cfRNAs) were extracted from 1 mL of plasma using the Plasma/Serum Circulating RNA and Exosomal Purification kit (Norgen). Purification was based on the use of Norgen's proprietary resin as the separation matrix. This kit is able to extract all sizes of circulating cfRNAs. The concentration of extracted cfRNAs was assessed using the Qubit RNA assay (Life Technologies).

Total cfRNA library (>50nt) was prepared with the SMARTer® Stranded Total RNA-Seq Kit–Pico. This kit removes ribosomal cDNAs after reverse transcription using a CRISPR/DASH method. We used Recombinant DNase I (TAKARA) to digest circulating DNAs. ERCC RNA Spike-In Control Mixes (Ambion) was added to the samples before library preparation, 1 µL per library at an appropriate concentration. RNA Clean and Concentrator-5 kit (Zymo) was used to obtain purified total RNA. More than 20 million reads of total cfRNA were sequenced on an Illumina platform for each library.

Potential contaminations in RNA extraction and library preparation were evaluated using two types of negative controls. 2 RNA samples were extracted from *E. coli* DH5α strain, using same kit for plasma cfRNA extraction. RNA-seq libraries of *E. coli* RNA samples, together with human brain RNA provided by SMARTer® Stranded Total RNA-Seq Kit, were constructed using same protocol for cfRNA library preparation.

Data processing

For RNA sequencing data, adapters and low-quality sequences in raw sequencing data were trimmed using cutadapt^[22] (version 2.3). GC oligos introduced in reverse transcription were also trimmed off, and reads shorter than 30 nt were discarded. We used STAR^[23] (version 2.5.3a_modified) for sequence mapping. The trimmed reads were mapped to ERCC's spike-in sequences, vector sequences in NCBI's UniVec database, and human rRNA sequences in refseq sequentially.

The remaining reads was mapped to the hg38 genome index built with the GENCODE^[24] v27 annotation. Circular RNA annotation was downloaded from circBase^[25]. Upstream 150bp and downstream 150bp sequences around the back spliced sites of circRNAs were concatenated to generate junction sequences, and circRNA sequences shorter than 100 bp were discarded. Reads unaligned to hg38 were mapped to circRNA junctions. Duplicates in the aligned reads were removed using Picard Tools MarkDuplicates (version 2.20.0). An aligned read pair was assigned to an RNA type if at least one of the mates overlapped with the corresponding genomic regions. In this way, the aligned reads were sequentially assigned to lncRNA, mRNA, snoRNA, snRNA, srpRNA and Y RNA with HTSeq^[26] package according to the GENCODE v27 annotation.

Count matrix for human genes was constructed using featureCounts^[27] v1.6.2 with the GENCODE v27 annotation. To avoid the impact of potential DNA contamination, only intron-spanning reads were considered.

Quality control

We filtered cfRNA-seq samples with multiple quality control criteria (Figure S1): 1) raw reads > 10 million; 2) clean reads (reads remained after trimming low quality and adaptor sequences) > 5 million; 3) aligned reads after duplicate removal (aligned to the human genome, hg38, and circRNA junctions) > 0.5 million; 4) For the clean reads, fraction of spike-in reads < 0.5 and ratio of rRNA reads < 0.5; 5) For genome aligned reads, ratio of mRNA and lncRNA reads > 0.2, ratio of unclassified reads < 0.3, and number of intron-spanning read pairs (defined as a read pair with a CIGAR string in which at least one mate contains "N" in the BAM files) > 100,000.

Differential analysis and functional enrichment analysis

We used quasi-likelihood method in edgeR^[28] package to identify differentially expressed genes and genera with significant abundance alteration ($|\log_2(\text{fold-change})| > 1$ and $\text{FDR} < 0.05$). We used this method to identify differential genes between cancer patients and healthy donors, as well as genes specific to one cancer type. For cancer type specific genes, previously reported gender related genes^[29] were excluded. KEGG pathway enrichment analysis of deregulated genes/RNAs were carried out using clusterProfiler^[30].

Data normalization

The count matrix of gene expression was normalized using the trimmed mean of M-values (TMM) method in edgeR. ANOVA was performed among different sample groups (HD and five cancer types) on discovery set using the quasi-likelihood method in edgeR, and the 25% most insignificant genes that stably expressed among different groups were considered as empirical control genes. The TMM normalized expression matrix was adjusted by the RUVg function in the RUVSeq^[31] package based on the identified control features.

Microbe data analysis

Unmapped reads (cleaned reads failed to align to human genome or circRNA junctions) were processed independently using a k-mer based pipeline and an alignment-based pipeline. In the first pipeline, unmapped reads were classified using kraken2^[32] with its standard database, which contains bacteria, archaea, virus and human sequences. In the alignment-based pipeline, using SortMeRNA^[33] (version 4.3.3), unmapped reads were annotated as either rRNA or non rRNA. rRNA reads were mapped to Silva database with bowtie2^[34]. Non rRNA reads were aligned to virus genome curated in kraken2's standard database. In both pipelines, counts at genus level were used for downstream analysis.

The same preprocessing and downstream analysis pipeline were applied to negative control samples (*E. coli* RNA-seq data was aligned to its reference genome NZ_CP025520.1 with bowtie2, instead of map to human rRNA, human genome and circRNA junctions). For reads coverage analysis of *Lawsonella clevelandensis* and HBV, reads unmapped to human sequences were mapped to their reference genomes (NZ_CP012390.1 and NC_003977.2, respectively).

Potential contaminations in genera detected by both kraken2 pipeline and bowtie2 pipeline (have at least 3 reads in at least 3 samples) were filtered prior to downstream analysis. We remove bacteria genera detected in at least 1 control samples (at least 3 reads), and virus genera detected in at least 1 *E. coli* control samples (at least 3 reads). Genera present in a previously reported common lab contamination list^[35], or genera that contain species with CPM > 10 in a published human skin microbiome dataset^[36] were removed. Virus genera that contains species with non-human eukaryotic host according to virushostdb^[37] were also excluded.

The genera with altered abundance were identified using edgeR. For cancer type specific microbes, genera with a prevalence lower than 20% in all sample groups were excluded. Counts at genus level were also normalized with TMM and RUVg, as we did for human gene expression.

Classification performance evaluation based on independent validation

We split the samples into a discovery set and an independent validation set in a 1:1 manner, stratified by cancer type, age, and gender. Esophageal cancer samples were all retained in the discovery set due to its small sample size.

In the discovery set, we use rank sum test to select 100 most significant features, and the more computationally intensive SURF method was applied to select 10 most important features. We used the ranksums functions in scipy^[38] for rank sum test. SURF was implemented in python package skrebate^[39]. To mitigate the impact of within-class heterogeneity, feature selection was embedded in the shuffle-split cross-validation process. The discovery set was randomly split into training and test sets in an 80%-20% manner; feature selection was performed on the training set and the model was evaluated on the test set. This procedure was repeated 100 times. This strategy was applied separately for gene expression and microbe abundance data, and 5 up regulated and 5 down regulated features were selected in each cross validation run.

Finally, 10 most frequently selected features were used for evaluate the binary classification performance on the validation set. A balanced version of random forest (implemented in python package imblearn^[40]) was used to handle the class-imbalance problem explicitly. AUROC and its confidence interval were calculated using R package pROC^[41].

For multiclass classification, we apply the same strategy in one vs. rest manner, that is selecting 10 up regulated features in each cross validation run, and finally used features selected with recurrence higher than 10 (in 100 cross-validation runs) for multiclass classification. In the final multi-classification model,

129, 117, and 145 features were selected from human gene expression, microbe genera abundance in kraken2 pipeline and bowtie2 pipeline, respectively.

Data availability

In addition to 35 datasets for liver cancer and 30 datasets for healthy donors we published previously (GSE142987)²¹, 230 raw FASTQ files supporting the findings of this study are available in the Gene Expression Omnibus (GEO: GSE174302).

For editors and reviewers:

The data in GSE174302 can be downloaded from the GEO with a secure token: cxmxyqcqlxctbst.

Results

Sequencing of cfRNAs captures signals of various long RNA species in plasma

Here we adapted a SMART based total RNA sequencing method (we called SMART-total) to profile plasma total cfRNAs. This technique was optimized for low input RNA sequencing, and robust for partially degraded RNA fragments. SMART-total was successfully applied to detect cfRNAs in plasma of pregnant women and cancer patients in previous studies^[17, 42, 43]. While most of these studies focus on human transcriptome, a study on plasma cfRNA of pregnant women suggested that microbial signal detected by this method can also provide some useful information^[14]. We applied SMART-total to a cohort of 295 plasma samples, the percentage of patients with early-stage cancer (stage I and II) range from 65% in stomach cancer to 86% in lung cancer (Table S1).

For low biomass metagenomic profiling, lab and kit contamination can lead to unreliable conclusion^[44]. Given the low concentration of both human and microbial cfRNAs in plasma, little contamination could have detrimental impacts on downstream analysis. To minimize impacts of potential microbe contamination introduced in sample collection, RNA extraction, library preparation, and sequencing, two *E. coli* samples and one human brain RNA sample were processed and sequenced following exactly the same procedure as plasma samples, serving as controls for contamination.

In addition to potential contaminations, misclassification of microbe derived reads may also render the result less interpretable. We carefully designed a computational pipeline to mitigate these problems (Figure 1A, see Methods). In brief, after removing human rRNA and other unwanted sequences, reads were aligned to human genome and circular RNA (circRNA) back spliced junctions to quantify human gene expression. Several quality control rules were applied to ensure data reliability, and there are 263 high quality samples remained for further analysis (Figure S1, Table S2). Unaligned reads were classified with kraken2^[32], an efficient but less stringent method based on k-mer contents; and a stringent but relatively computationally intensive method based on bowtie2 alignment^[34]. Since the majority of microbial reads are rRNA, we only mapped microbial rRNA reads against Silva database^[45] to reduce the

computational burden. The rest non-rRNA reads were aligned to viral genomes. From the resulting microbial profile, we filtered away genera that were found in our control samples (Table S3), previously reported common lab contaminations^[35] or skin bacteria^[36], which are often regarded as potential sources of contamination^[46]. Some suspicious viral genera with non-human eukaryotic hosts^[37] were also excluded (Table S3).

Using this computational pipeline, the majority of cleaned reads were mapped to human genome (79.36% on average) and back-spliced junctions of circular RNA (1.24% on average). In the remaining reads, 10.18% were annotated as non-human rRNA, and 2.06% were further assigned to microbial genomes by kraken2 (Figure 1B, Table S4).

Consistent with intracellular long RNA profile, mRNA and lncRNA are the most abundant human RNA species captured in SMART-total library (Figure 2A). Some house-keeping genes, such as ACTB, TUBB1, PTMA, and noncoding RNAs, such as srpRNA (RN7SL2), are highly abundant in plasma of both cancer patients and healthy donors (Figure S2A). For these transcripts, the coverages are uniformly distributed along the full-length transcripts in samples from different clinical centers (Figure 2B). Previous studies demonstrated mRNAs mainly exist as short fragments up to several hundred nucleotides^[47]. This uniform coverage indicates that at least for these most abundant transcripts, such naturally occurred fragmentation process does not have strong sequence bias. Meanwhile, a sharp boundary of read coverage at exon-intron junctions further demonstrated that there were minimal genomic DNA contaminations in our sequencing libraries (Figure 2C).

As for microbe derived reads, the most abundant phylum is *Proteobacteria*, followed by *Firmicutes* and *Actinobacteria* (Figure 2D), which resembles previous reports for microbe derived cfDNA and cfRNA in plasma^[15,17,48–50]. Consistent with previous studies^[51], the order *Caudovirales*, which includes tailed bacteriophages, makes up the majority (the median fraction is higher than 95%) of reads assigned to viruses by kraken2.

We also investigated the read coverage for detected microbes by aligning non-human reads to their genomes. As expected, for bacteria, most of the RNA-seq signals agree with the previous notion that most microbial reads are from rRNA, as shown in Figure 2E for *Lawsonella clevelandensis* (a pathogen reported to induce abscess^[52]) as an example. The RNA-seq signals for viruses are also consistent with their genome annotations. For instance, in a representative coverage of HBV genome (Figure 2F), the reads coverage of gene X agrees well with its annotated boundary.

cfRNA profile alterations in patients are cancer relevant

To investigate the biological relevance of plasma cfRNAs in cancer patients, the enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways of human genes differentially expressed in cancer patients (Table S5) were identified (Table 1). Enriched pathways of up-regulated genes include ECM-receptor interaction and neutrophil extracellular trap (NET), which have been recognized to promote metastasis⁵³. Down-regulated cfRNAs are highly enriched in pathways mainly related to ribosome

biogenesis. Down-regulation of translation-related pathways was previously reported in tumor educated platelets (TEPs)^[4], indicating that translational events might be globally suppressed in the blood milieu of cancer patients. More interestingly, multiple immune related pathways (PD-1 checkpoint, T cell differentiation, NOD-like receptor signaling, cytosolic DNA-sensing and NF-κB signaling) are downregulated in cancer patients, depicting their suppressed immune status. These findings suggest that signals related to tumor and tumor microenvironment can be identified by cfRNA-seq. For comparisons among different cancer types and HDs, similar patterns were also observed (Figure S3).

For microbial cfRNAs, we found that the plasma abundance of multiple viral genera, include *Lymphocryptovirus*, *Mastadenovirus*, *Roseolovirus*, several genera of torque teno viruses (TTVs), and *Orthohepadnavirus* are significantly higher in cancer patients (Figure 3A). This result is supported by both pipelines (Table S5). The viral load of two prevalent genera, *Alphatorquevirus* and *Orthohepadnavirus* are associated with liver cancer (Figure 3B). TTVs are highly prevalent viruses even in healthy population, and are not considered as pathogens of a specific disease, but associations between TTV and liver diseases have been widely reported⁵⁴. Higher TTV abundance is also associated with suppressed immune status, and has been utilized as an indicator of immunosuppression after organ transplantation⁵⁴⁻⁵⁶. Its enrichment in cancer patients is concordant with the down regulation of immune pathways we found in human cfRNAs. The association between liver cancer and *Orthohepadnavirus*, a genus to which HBV belongs, is expected, as 60% liver cancer patients in this study have a history of HBV induced chronic hepatitis (Table S1). Other viral genera that significantly altered in liver cancer were also shown (Figure 3C).

Cancer type specificity of human and microbial cfRNAs

We further investigated the cancer specificity of human and microbial cfRNAs by comparing each cancer type to the remaining ones (Table S6). Significant human genes and microbial genera with highest fold changes are listed for each cancer type respectively (Figure 4A, Figure 4B). Several circular RNAs are more abundant in plasma of stomach cancer and colorectal cancer patients. In plasma of liver cancer patients, there is a dramatic increase in the expression level of CP gene, which encodes plasma protein ceruloplasmin. CP is a liver specific gene, as ceruloplasmin is synthesized in hepatocytes, and then secreted into blood⁵⁷. This indicates that some intracellular RNA contents in liver tumor sites can be directly reflected in plasma. CLEC4E and IFIT1B, which were identified as lung cancer and esophageal cancer specific respectively (Figure 4A), are related to immune regulation^{58,59}. Colorectal cancer is associated with higher abundance of *Mycoplasma* and *Acholeplasma* derived cfRNAs. Consistently, *Orthohepadnavirus* and TTVs are again identified as liver cancer specific. Taken together, our analysis suggests the abundance of human and microbial cfRNAs in plasma both have cancer type specificity.

Evaluating cancer detection capacity of human and microbial cfRNAs

We used a machine learning method (Figure S4, see Methods) to evaluate the capacity of the plasma cfRNA abundance in distinguishing cancer patients from HDs. Samples were split into a discovery set

and a validation set, stratified by genders, ages, and sample sources (Table S7). All of the esophageal cancer samples were reserved in the discovery set for its relatively small sample size.

We used results of both k-mer based and alignment based pipelines for machine learning analysis. As kraken2 assigns reads of different samples independently to certain taxonomy in a consistent manner, its higher false positive rate should have little influence on estimation of classification performance.

For both human and microbial reads, we normalized the data, and performed batch correction with RUVg³¹ (Figure S5, see Methods). We performed feature selection, and fit random forest classifiers with 10 selected features for performance evaluation (See Methods).

Cross validation performance on the discovery set for binary classification of cancer patients vs. HDs was illustrated (Figure 5A-B). The average cross-validation AUROC scores of human cfRNAs were ~0.8-0.95 (Random choices would provide AUROC scores of around 0.5). Microbial cfRNAs quantified by k-mer based pipeline could also achieve AUROC of ~0.75-0.95. We evaluated the performance of 10 recurrently selected human and microbe features on the validation set. Human and microbial cfRNAs both showed capacity of separating cancer patients from HDs. Combining human and microbe features could improve the prediction performance for cancer vs. HD classification, and achieved an AUROC (area under ROC curve) of 0.931 (Figure 5C). Comparable performances were observed when using alignment based method in most cases (Figure S6).

Microbial cfRNAs remarkably enhance the classification of multiple cancer types

Given that the cfRNA profile could distinguish cancer patients from HDs, we further assessed the feasibility of using cfRNAs for classifying cancer patients with different primary tumor locations. In the discovery set, we first selected features capable of distinguishing each cancer type from the remaining ones, then took the union of these features to fit a random forest model for multiclass classification (Table S8).

We used leave-one-out cross-validation to evaluate performance of these features on the discovery set (Figure 5D-E). On the validation set, using human cfRNA features, an average accuracy of 51.3% was achieved, 29.2% of colorectal cancer, 41.2% of stomach cancer, 78.6% of liver cancer and 56.3% of lung cancer patients were correctly classified (Figure 5D). Using microbial cfRNAs features in k-mer based pipeline, 45.8% of colorectal cancer, 88.2% of stomach cancer, 60.7% of liver cancer and 31.3% of lung cancer patients were correctly classified (Figure 5E). This performance can be further improved when microbial features were combined with human features, and an average accuracy of 62.8% was achieved on validation set using combined features, led to 11.5% improvement compared to using human feature alone (Figure 5F, Figure S7). The multiclass classification performance was marginally worse for alignment-based pipeline (Figure S6), but still much better than random guess, as for each cancer type, random assign each sample to one of five cancer types should give top 1 accuracy of 20% and top 2 accuracy of 40%. Taken together, the human and microbial fraction in plasma cfRNAs both provide tumor site specific information, and is predictive for the primary location of tumors.

Discussion

In this study, we sequenced cfRNAs in a cohort of patients with five major types of highly malignant cancer. We demonstrated that there are biological relevant differences between cfRNA of healthy donors and cancer patients. Cancer type specific signals could be identified in both human and microbial cfRNAs, and these signals could be utilized to detect and classify multiple cancer types, including early-stage cases. Some of liver cancer specific signals are rightly interpretable, such as tissue specific genes (CP gene) and well-known viruses (HBVs and TTVs). circRNAs were reported to have tissue specificity^[60], and purposed as plasma cancer biomarker^[5]. In this study, several plasma circRNAs also show some cancer type specificity, especially for colorectal and stomach cancer. *Mycoplasma* and *Acholeplasma* are identified as colorectal cancer specific in our cfRNA profiles. The relevance between *Mycoplasma* infection and cancers was previously reported^[61, 62], *Acholeplasma* was also reported to be more abundant in gut microbiome of colon cancer patients^[63]. The remaining ones can be potentially explained by secondary signals that reflect interactions between tumor and certain blood components such as immune cells and platelets, or some uncharacterized interactions between human and microbes.

The existence of microbe derived plasma nucleic acids in donors without sepsis has been independently demonstrated by multiple studies. In typical bioinformatic analysis, reads that cannot be aligned to human genome are discarded. Our work suggests these data can be further exploited, and provide useful information for microbial profiling in plasma. Several studies suggested that the human virome at different body sites, including plasma, has an unexpected diversity^[16, 51], and our current knowledges of human virome are limited to species that could cause serious clinical consequences. Our work highlights the feasibility of discovering clinically relevant but understudied viruses from high throughput sequencing data.

Although RNA is more prone to degradation, RNA-seq does have some favorable properties compared to DNA-seq in detecting microbial signals. Obviously, DNA-seq cannot detect RNA viruses. In addition, it has been reported that the microbe derived cfDNA only makes up a small fraction (lower than 0.5% in some cases) in human cfDNA pool^[15, 16, 64]. The genome of bacteria and viruses are much more compact than human, and a larger fraction of their genome sequence is transcribed into RNAs. That indicates if a mixture of human cells and bacteria is sequenced by DNA-seq and RNA-seq to same depth, microbial reads should make up a larger fraction (around 10% on average in our study) in RNA-seq library, and microbial signal can be captured more efficiently.

Confounding effect is a major obstacle for discovering reliable biomarkers from high throughput data. In our cohort design, samples were collected from different clinical centers, and genders for some cancer types, like liver cancer, were not well balanced. We attempted to mitigate the problems computationally by using RUVg to remove these unwanted variations. Our analysis provided clues for the clinical relevance of microbe derived cfRNAs, but further study with a larger, carefully designed cohort is still required for clinical application.

Conclusion

Taken together, we provide evidence for the clinical relevance of human and microbe derived plasma cfRNAs and their potential utilities in cancer detection, and determination of tumor sites. Combining microbe and human derived cfRNAs might shed light on the development of novel strategies for early detection and classification of multiple cancer types.

Abbreviations

cfRNAs: cell free RNAs; ROC: receiver operating characteristic; cfDNA: cell free DNA; EVs: extracellular vesicles; HBV: hepatitis B virus; HPV: human papillomavirus; HDs: healthy donors; KEGG: Kyoto Encyclopedia of Genes and Genomes; NET: neutrophil extracellular trap; ECM: extracellular matrix; TEPs: tumor educated platelets; TTVs: torque teno viruses; RUVg: remove unwanted variation using control genes; AUROC: area under receiver operating characteristic curve.

Declarations

Ethical Approval and Consent to participate

The study was approved by the local institutional research ethics committees. Informed consent was obtained from all patients.

Consent for publication

Not applicable.

Availability of data and materials

Raw FASTQ files supporting the findings of this study are available in the Gene Expression Omnibus (GEO: GSE174302). *For editors and reviewers* The data in GSE174302 can be downloaded from the GEO with a secure token: cmxycqelxctbst.

Competing interests

The authors have declared that no conflict of interest exists.

Funding

This work was supported by National Natural Science Foundation of China (31771461, 81972798, 81373067, 81773140, 81902384), National Key Research and Development Plan of China (2017YFA0505803, 2017YFC0908401, 2019YFC1315700), National Science and Technology Major Project of China (2018ZX10723204, 2018ZX10302205) Tsinghua University Initiative Scientific Research Program (2021Z99CFY022), Tsinghua-Foshan Innovation Special Fund, and Fok Ying-Tong Education Foundation. This study was also supported by Interdisciplinary Clinical Research Project of Peking

Authors' contributions

S.C., S.W., Z.J.L, Z.Z.X. and P.W. conceived and designed the project. S.W., S.X., P.B., and Y.M. performed the experiments. Y.J. and Y.T. processed the data and completed the bioinformatically analyses. Sample and clinical information were collected by S.C., S.Z., H.C., Y.L., F.X., C.X., J.Y., P.W. All authors contributed to the final version of the manuscript.

References

1. Abbosh, C. *et al.* Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. *Nature* **545**, 446-451, doi:10.1038/nature22364 (2017).
2. Shen, S. Y. *et al.* Sensitive tumour detection and classification using plasma cell-free DNA methylomes. *Nature* **563**, 579-583, doi:10.1038/s41586-018-0703-0 (2018).
3. Cristiano, S. *et al.* Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature* **570**, 385-389, doi:10.1038/s41586-019-1272-6 (2019).
4. Best, Myron G. *et al.* RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics. *Cancer Cell* **28**, 666-676, doi:<https://doi.org/10.1016/j.ccell.2015.09.018> (2015).
5. Li, Y. *et al.* Circular RNA is enriched and stable in exosomes: a promising biomarker for cancer diagnosis. *Cell Research* **25**, 981-984, doi:10.1038/cr.2015.82 (2015).
6. Tan, C. *et al.* Noncoding RNAs Serve as Diagnosis and Prognosis Biomarkers for Hepatocellular Carcinoma. *Clinical Chemistry* **65**, 905-915, doi:10.1373/clinchem.2018.301150 (2019).
7. Arbuthnot, P. & Kew, M. Hepatitis B virus and hepatocellular carcinoma. *Int J Exp Pathol* **82**, 77-100, doi:10.1111/j.1365-2613.2001.iep0082-0077-x (2001).
8. Burd, E. M. Human papillomavirus and cervical cancer. *Clin Microbiol Rev* **16**, 1-17, doi:10.1128/CMR.16.1.1-17.2003 (2003).
9. Polk, D. B. & Peek, R. M. Helicobacter pylori: gastric cancer and beyond. *Nature Reviews Cancer* **10**, 403-414, doi:10.1038/nrc2857 (2010).
10. Han, Y. W. Fusobacterium nucleatum: a commensal-turned pathogen. *Current Opinion in Microbiology* **23**, 141-147, doi:<https://doi.org/10.1016/j.mib.2014.11.013> (2015).
11. Riquelme, E. *et al.* Tumor Microbiome Diversity and Composition Influence Pancreatic Cancer Outcomes. *Cell* **178**, 795-806.e712, doi:<https://doi.org/10.1016/j.cell.2019.07.008> (2019).
12. Gosiewski, T. *et al.* Comprehensive detection and identification of bacterial DNA in the blood of patients with sepsis and healthy volunteers using next-generation sequencing method - the observation of DNAemia. *European Journal of Clinical Microbiology & Infectious Diseases* **36**, 329-336, doi:10.1007/s10096-016-2805-7 (2017).

13. Blauwkamp, T. A. *et al.* Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease. *Nature Microbiology* **4**, 663-674, doi:10.1038/s41564-018-0349-6 (2019).
14. Potgieter, M., Bester, J., Kell, D. B. & Pretorius, E. The dormant blood microbiome in chronic, inflammatory diseases. *FEMS Microbiology Reviews* **39**, 567-591, doi:10.1093/femsre/fuv013 (2015).
15. Zozaya-Valdés, E. *et al.* Detection of cell-free microbial DNA using a contaminant-controlled analysis framework. *Genome Biology* **22**, 187, doi:10.1186/s13059-021-02401-3 (2021).
16. Kowarsky, M. *et al.* Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA. *Proceedings of the National Academy of Sciences* **114**, 9623, doi:10.1073/pnas.1707009114 (2017).
17. Pan, W. *et al.* Simultaneously Monitoring Immune Response and Microbial Infections during Pregnancy through Plasma cfRNA Sequencing. *Clinical Chemistry* **63**, 1695-1704, doi:10.1373/clinchem.2017.273888 (2017).
18. Luo, Z. *et al.* CRIg+ Macrophages Prevent Gut Microbial DNA Containing Extracellular Vesicle Induced Tissue Inflammation and Insulin Resistance. *Gastroenterology* **160**, 863-874, doi:10.1053/j.gastro.2020.10.042 (2021).
19. Poore, G. D. *et al.* Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* **579**, 567-574, doi:10.1038/s41586-020-2095-1 (2020).
20. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2015. *CA: A Cancer Journal for Clinicians* **65**, 5-29, doi:<https://doi.org/10.3322/caac.21254> (2015).
21. Zhu, Y. M. *et al.* Integrative analysis of long extracellular RNAs reveals a detection panel of noncoding RNAs for liver cancer. *Theranostics* **11**, 181-193, doi:10.7150/thno.48206 (2021).
22. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*; Vol 17, No 1: Next Generation Sequencing Data Analysis DO - 10.14806/ej.17.1.200 (2011).
23. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 %J Bioinformatics (2012).
24. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-1774, doi:10.1101/gr.135350.111 (2012).
25. Glažar, P., Papavasileiou, P. & Rajewsky, N. circBase: a database for circular RNAs. *RNA* **20**, 1666-1670, doi:10.1261/rna.043687.113 (2014).
26. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)* **31**, 166-169, doi:10.1093/bioinformatics/btu638 (2015).
27. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930, doi:10.1093/bioinformatics/btt656 (2013).

28. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140, doi:10.1093/bioinformatics/btp616 %J Bioinformatics (2009).
29. Shi, M.-W. *et al.* SAGD: a comprehensive sex-associated gene database from transcriptomes. *Nucleic Acids Research* **47**, D835-D840, doi:10.1093/nar/gky1040 (2018).
30. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284-287, doi:10.1089/omi.2011.0118 (2012).
31. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology* **32**, 896-902, doi:10.1038/nbt.2931 (2014).
32. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biology* **20**, 257, doi:10.1186/s13059-019-1891-0 (2019).
33. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**, 3211-3217, doi:10.1093/bioinformatics/bts611 (2012).
34. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).
35. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology* **12**, 87, doi:10.1186/s12915-014-0087-z (2014).
36. Oh, J., Byrd, Allyson L., Park, M., Kong, Heidi H. & Segre, Julia A. Temporal Stability of the Human Skin Microbiome. *Cell* **165**, 854-866, doi:<https://doi.org/10.1016/j.cell.2016.04.008> (2016).
37. Mihara, T. *et al.* Linking Virus Genomes with Host Taxonomy. *Viruses* **8**, doi:10.3390/v8030066 (2016).
38. Virtanen, P. *et al.* SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **17**, 261-272 (2020).
39. Urbanowicz, R. J., Olson, R. S., Schmitt, P., Meeker, M. & Moore, J. H. Benchmarking Relief-Based Feature Selection Methods. (2017).
40. Lemaître, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research* **18**, 1-5 (2017).
41. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77-77, doi:10.1186/1471-2105-12-77 (2011).
42. Ngo Thuy, T. M. *et al.* Noninvasive blood tests for fetal development predict gestational age and preterm delivery. *Science* **360**, 1133-1136, doi:10.1126/science.aar3819 (2018).
43. Yu, S. *et al.* Plasma extracellular vesicle long RNA profiling identifies a diagnostic signature for the detection of pancreatic ductal adenocarcinoma. *Gut* **69**, 540, doi:10.1136/gutjnl-2019-318860 (2020).
44. Eisenhofer, R. *et al.* Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends in Microbiology* **27**, 105-117, doi:<https://doi.org/10.1016/j.tim.2018.11.003> (2019).

45. Yilmaz, P. *et al.* The SILVA and “All-species Living Tree Project (LTP)” taxonomic frameworks. *Nucleic Acids Research* **42**, D643-D648, doi:10.1093/nar/gkt1209 (2014).
46. Schierwagen, R. *et al.* Trust is good, control is better: technical considerations in blood microbiome analysis. *Gut* **69**, 1362, doi:10.1136/gutjnl-2019-319123 (2020).
47. Larson, M. H. *et al.* A comprehensive characterization of the cell-free transcriptome reveals tissue- and subtype-specific biomarkers for cancer detection. *Nature Communications* **12**, 2357, doi:10.1038/s41467-021-22444-1 (2021).
48. Yao, J., Wu, D. C., Nottingham, R. M. & Lambowitz, A. M. Identification of protein-protected mRNA fragments and structured excised intron RNAs in human plasma by TGIRT-seq peak calling. *eLife* **9**, e60743, doi:10.7554/eLife.60743 (2020).
49. Païssé, S. *et al.* Comprehensive description of blood microbiome from healthy donors assessed by 16S targeted metagenomic sequencing. *Transfusion* **56**, 1138-1147, doi:<https://doi.org/10.1111/trf.13477> (2016).
50. Lelouvier, B. *et al.* Changes in blood microbiota profiles associated with liver fibrosis in obese patients: A pilot analysis. *Hepatology* **64**, 2015-2027, doi:<https://doi.org/10.1002/hep.28829> (2016).
51. Liang, G. & Bushman, F. D. The human virome: assembly, composition and host interactions. *Nature Reviews Microbiology* **19**, 514-527, doi:10.1038/s41579-021-00536-5 (2021).
52. Goldenberger, D. *et al.* Emerging anaerobic and partially acid-fast *Lawsonella clevelandensis*: extended characterization by antimicrobial susceptibility testing and whole genome sequencing. *Clinical Microbiology and Infection* **25**, 1447-1448, doi:10.1016/j.cmi.2019.07.008 (2019).
53. Xiao, Y. *et al.* Cathepsin C promotes breast cancer lung metastasis by modulating neutrophil infiltration and neutrophil extracellular trap formation. *Cancer Cell* **39**, 423-437.e427, doi:<https://doi.org/10.1016/j.ccell.2020.12.012> (2021).
54. Mrzljak, A. & Vilibic-Cavlek, T. Torque teno virus in liver diseases and after liver transplantation. *World J Transplant* **10**, 291-296, doi:10.5500/wjt.v10.i11.291 (2020).
55. Jaksch, P. *et al.* Torque Teno Virus as a Novel Biomarker Targeting the Efficacy of Immunosuppression After Lung Transplantation. *The Journal of Infectious Diseases* **218**, 1922-1928, doi:10.1093/infdis/jiy452 (2018).
56. Spandole, S., Cimponeriu, D., Berca, L. M. & Mihăescu, G. Human anelloviruses: an update of molecular, epidemiological and clinical aspects. *Archives of Virology* **160**, 893-908, doi:10.1007/s00705-015-2363-9 (2015).
57. Tao, T. Y. & Gitlin, J. D. Hepatic copper metabolism: Insights from genetic disease. *Hepatology* **37**, 1241-1247, doi:<https://doi.org/10.1053/jhep.2003.50281> (2003).
58. Patin, E. C., Orr, S. J. & Schaible, U. E. Macrophage Inducible C-Type Lectin As a Multifunctional Player in Immunity. *Front Immunol* **8**, 861-861, doi:10.3389/fimmu.2017.00861 (2017).
59. Daugherty, M. D., Schaller, A. M., Geballe, A. P. & Malik, H. S. Evolution-guided functional analyses reveal diverse antiviral specificities encoded by IFIT1 genes in mammals. *eLife* **5**, e14228, doi:10.7554/eLife.14228 (2016).

60. Xia, S. *et al.* Comprehensive characterization of tissue-specific circular RNAs in the human and mouse genomes. *Briefings in Bioinformatics* **18**, 984-992, doi:10.1093/bib/bbw081 (2017).

61. Huang, S., Li, J. Y., Wu, J., Meng, L. & Shou, C. C. Mycoplasma infections and different human carcinomas. *World J Gastroenterol* **7**, 266-269, doi:10.3748/wjg.v7.i2.266 (2001).

62. Zella, D. *et al.* Mycoplasma promotes malignant transformation in vivo, and its DnaK, a bacterial chaperone protein, has broad oncogenic properties. *Proceedings of the National Academy of Sciences* **115**, E12005, doi:10.1073/pnas.1815660115 (2018).

63. Shoji, M. *et al.* Characteristics of the gut microbiome profile in obese patients with colorectal cancer. *JGH Open* **5**, 498-507, doi:10.1002/jgh3.12529 (2021).

64. Xiao, Q. *et al.* Alterations of circulating bacterial DNA in colorectal cancer and adenoma: A proof-of-concept study. *Cancer Letters* **499**, 201-208, doi:<https://doi.org/10.1016/j.canlet.2020.11.030> (2021).

Table

Due to technical limitations, table 1 tiff is only available as a download in the Supplemental Files section.

Figures

Figure 1

Pipeline for cfRNA sequencing data processing A. The bioinformatic pipeline for plasma cfRNA sequencing data processing. After adapter trimming, spike in, potential vector contaminations and human rRNA sequences were removed. Cleaned reads were aligned to human genome and circular RNA

back spliced junctions. Unmapped reads were classified with a k-mer based and an alignment based pipeline. Genera detected by both pipelines were used for downstream analysis. Potential contaminations (known common lab contaminants, genera detected in control samples, skin microbes, and suspicious viral genera) were excluded. See Methods section for detail. B. Average fractions of different cfRNA components in cleaned reads. Microbe-rRNA refers to reads annotated to rRNA by sortmerna. Microbe-others refers to non-rRNA reads that were assigned to microbial genomes by kraken2.

Figure 2

Human genes and microbial signal revealed by cfRNA-seq A. The number of detected human transcripts (CPM > 2) of different RNA types and their relative abundances. CPM: counts per million. B. Representative coverages for ACTB and TUBB1 in healthy donors from three clinical centers (Samples HD-1, HD-2 and HD-3 are provided by PKU, ShH-1 and SWU, respectively). C. Metagene plot for read coverage around 5' exon boundaries and 3' exon boundaries. The mean coverage of 100 nt around exon boundaries for exons with read coverage > 3 were shown. D. Relative abundance of reads assigned to different phyla by kraken2. E. Representative reads coverages of Lawsonella clevelandensis 16S and 23S rRNA in the healthy donors from three clinical centers. F. A representative reads coverage on HBV genome in cfRNA of a patient with liver cancer.

Figure 3

Figure 3

Biological relevance of alteration in microbial cfRNA profile A. Example genera with significantly altered abundance in cancer patients when compared to HDs. FC: fold change. FC and FDR were calculated using result of alignment based method, and labeled genera were supported by both pipelines. B. Abundance of Alphatorquevirus and Othohepavirus in alignment based pipeline across different samples ranked in descending order, colors indicate different sample groups. C. Virus genera with significant abundance alteration ($FDR < 0.05$ and $\log_2\text{FoldChange} > 1$) in liver cancer patients when compared to HDs.

Figure 4

Cancer type specificity of human and microbe derived cfRNAs A. Significantly upregulated human gene expression ($FDR < 0.05$) in each cancer type when compared to the remaining ones. 5 features with highest fold change were shown. Sizes of the circle indicate the relative abundances (scaled to 0-1) across different cancers, colors of the circles indicate P values. B. Microbial genera with significant higher abundance in one cancer type vs. rest comparisons. For significant genera ($FDR < 0.05$ in both kraken2 and bowtie2's result) of each cancer type with prevalence higher than 20% in at least one group of patients, the top 5 genera with highest fold change were shown. Sizes and colors of the circles indicate the relative abundances (bowtie2's result, scaled to 0-1) and P values, respectively.

Figure 5

Detection and classification of cancer using human and microbial cfRNAs A-B. Cross-validation performance on the discovery set using abundance of human genes and microbial genus (kraken2's results) for the binary classification between cancer patients and healthy donors. On the discovery set, an 80% random subset of samples were used for feature selection and model training, and the performance was evaluated on the remaining 20% samples. This procedure is repeated for 100 times, yield 100 AUROC scores for each binary classification, as shown in the box plot. C. AUROC of binary classification on the validation set, using abundance of human genes, microbial genera, and combining both features. Error bars indicate 95% confidence intervals. D-E. Confusion matrix of multiclass classification using human genes (D) and microbial abundance (E) in kraken2 pipeline. Rows indicate true labels, columns indicate predicted labels. Left panel: leave-one-out performance on the discovery set; right panel: performance on the validation set. *as esophageal cancer samples are all reserved in discovery set, the classification performance of esophageal cancer samples on validation set is not applicable. F. Accuracy of top 1 and top 2 predictions on the discovery set (left) and validation set (right), using human, microbial and combined features.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementarymaterial.zip](#)
- [Tabel1.tif](#)