RESEARCH ARTICLE

Figures and figure supplements

Cancer type classification using plasma cell-free RNAs derived from human and microbes

Shanwen Chen *et al*

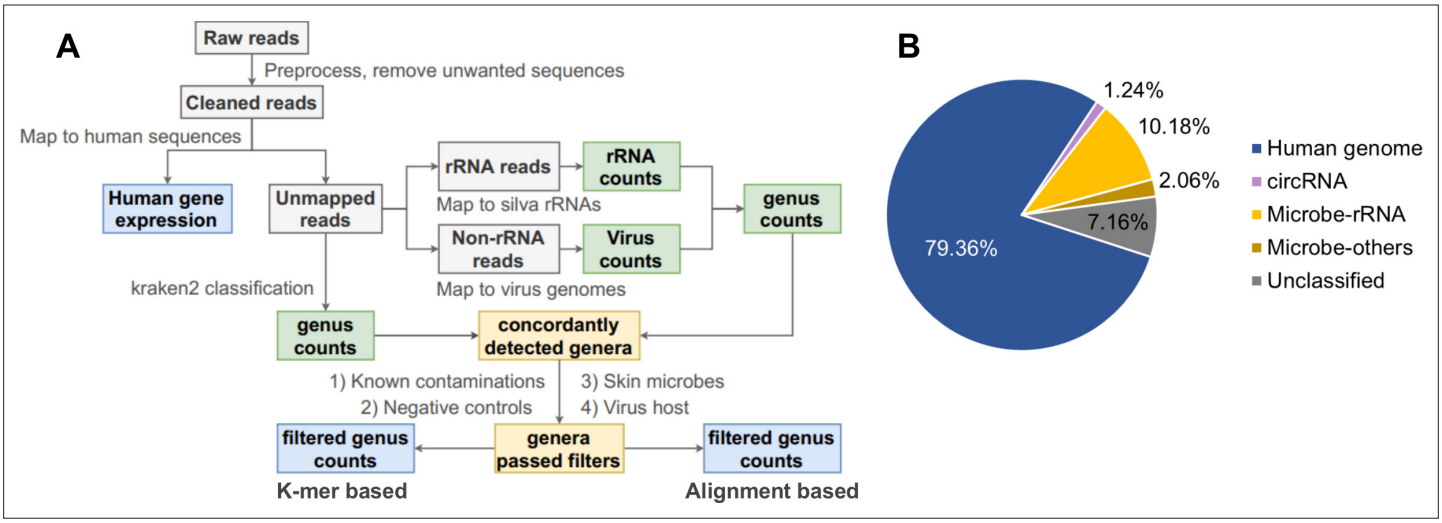

Figure 1. Pipeline for cell-free RNA (cfRNA) sequencing data processing. (**A**) The bioinformatic pipeline for plasma cfRNA sequencing data processing. After adapter trimming, spike in, potential vector contaminations, and human rRNA sequences were removed. Cleaned reads were aligned to the human genome and circular RNA back-spliced junctions. Unmapped reads were classified with a k-mer-based pipeline and an alignment-based pipeline. Genera detected by both pipelines were used for downstream analysis. Potential contaminations (known common laboratory contaminants, genera detected in control samples, skin microbes, and suspicious viral genera) were excluded. See the Materials and methods section for details. (**B**) Average fractions of different cfRNA components in cleaned reads. Microbe-rRNA refers to reads annotated to rRNA. Microbe-others refers to non-rRNA reads that were assigned to microbial genomes by kraken2.

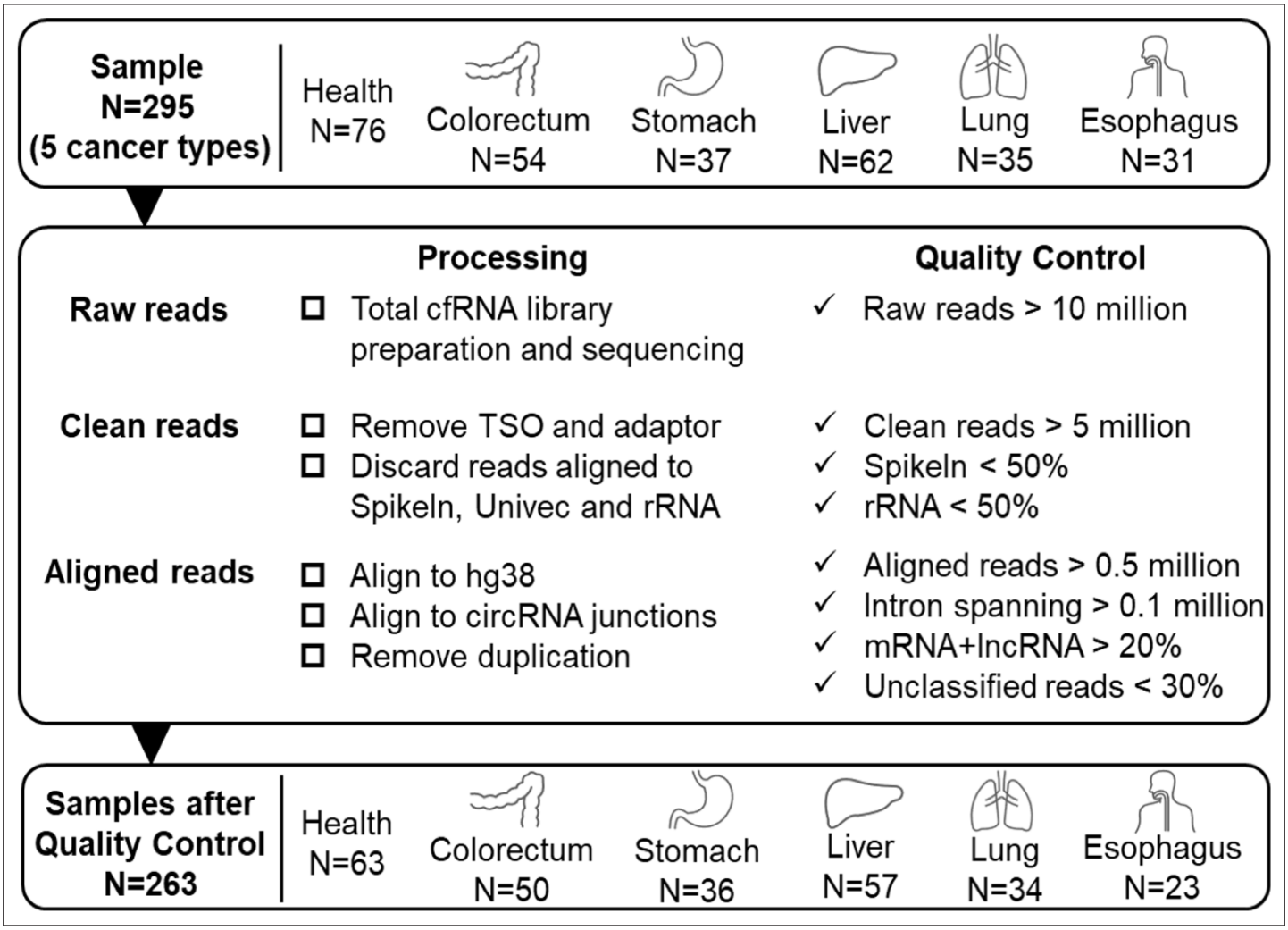

Figure 1—figure supplement 1. Quality control of sequencing data. We used 295 samples of 5 cancer patients and healthy donors to discover potential RNA biomarkers. 263 samples passed the following quality control criteria: (1) Raw reads: at least 10 million. (2) Clean reads: at least 5 million. Spike-in sequence: <50%. rRNA sequence: <50%. (3) Aligned reads (reads mapped to the human genome or circRNA junctions): at least 0.5 million; Intron-spanning reads: at least 0.1 million; reads assigned to mRNA and lncRNA: at least 20%; unclassified reads (reads cannot be assigned to annotated exon, intron, antisense of exons, promoter, enhancer, or repeats): less than 30%.

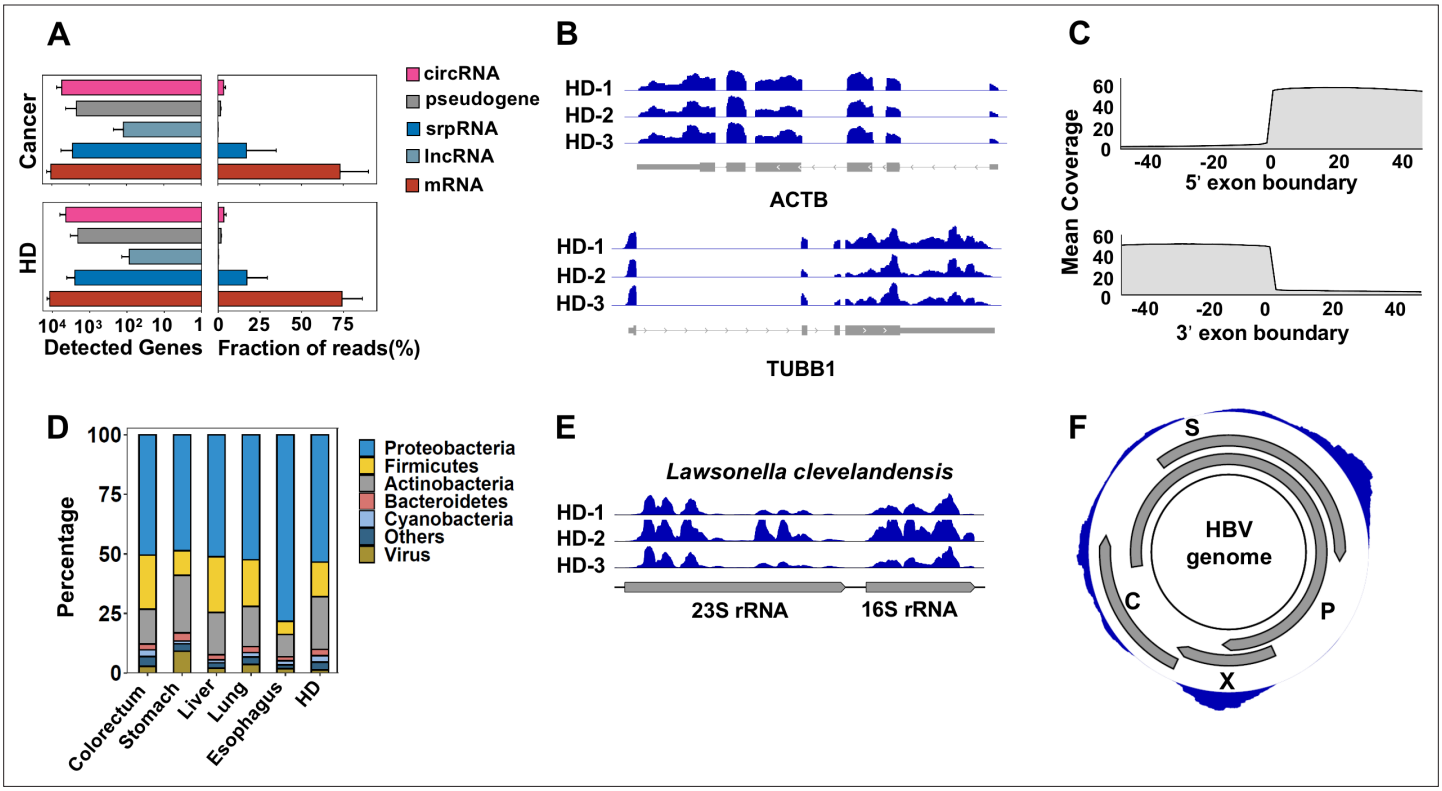

Figure 2. Human genes and microbial signals revealed by cell-free RNA (cfRNA)-seq. (**A**) The number of detected human transcripts (counts per million >2) of different RNA types and their relative abundances. (**B**). Representative coverages for ACTB and TUBB1 in healthy donors (HDs) from three clinical 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 is shown. (**D**). Relative abundance of reads assigned to different phyla by kraken2. (**E**). Representative read coverage of *Lawsonella clevelandensis* 16S and 23S rRNA in healthy donors from three clinical centers. (**F**). A representative read coverage on the HBV genome in cfRNA of a patient with liver cancer.

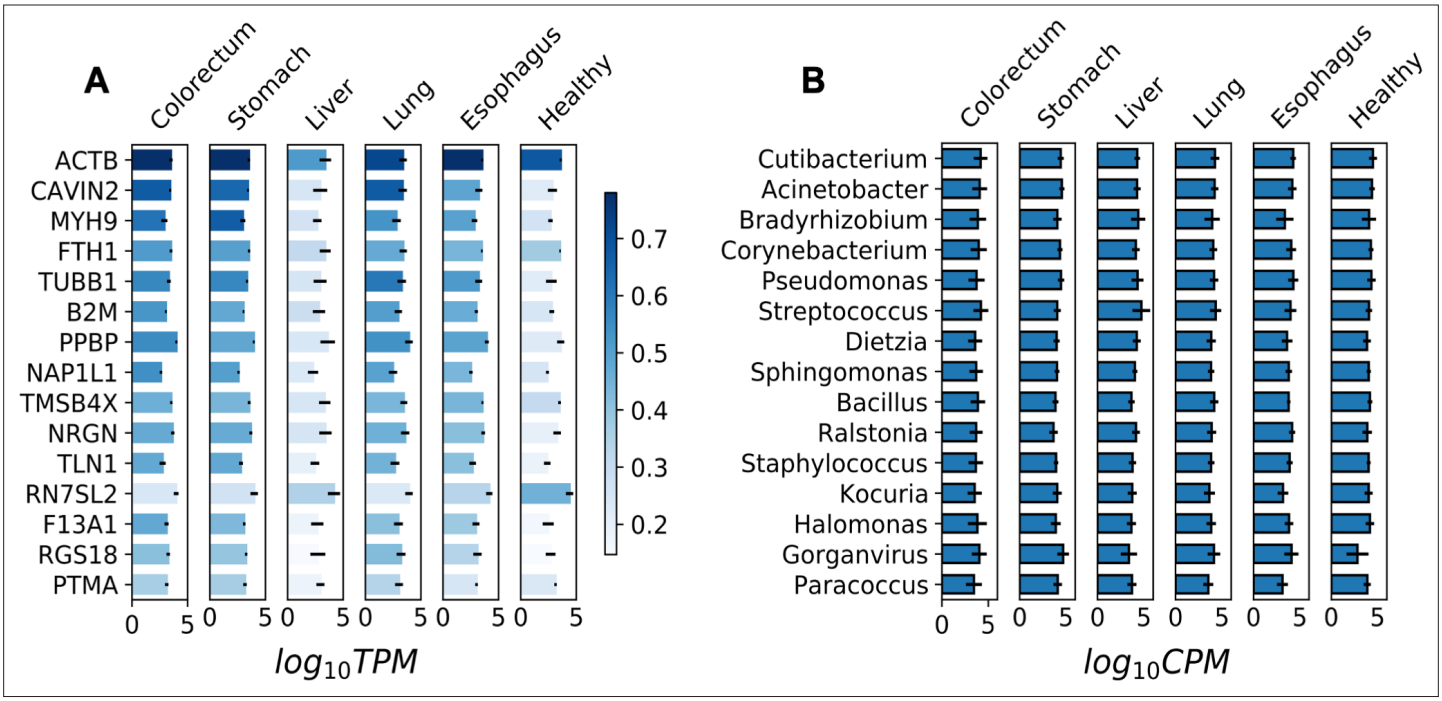

Figure 2—figure supplement 1. Most abundant human genes and microbial genera in plasma cell-free (cfRNA) libraries. (**A**) The abundance (log10TPM) of the 15 most abundant human genes in different sample groups. TPM: transcripts per million. (**B**) The abundance (log10CPM) of the 15 most abundant genera in different sample groups. CPM: counts per million.

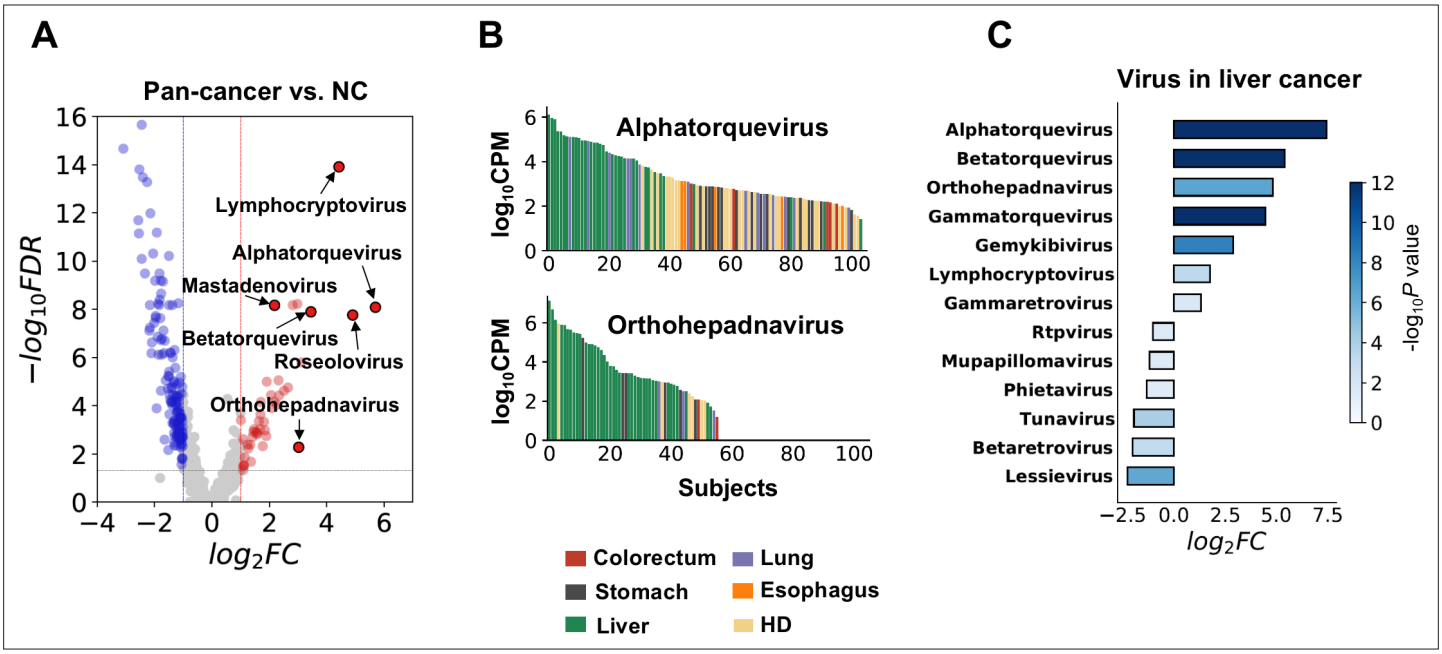

Figure 3. Biological relevance of alterations in the microbial cell-free RNA (cfRNA) profile. (**A**) Example genera with significantly altered abundance in cancer patients when compared to healthy donors (HDs). FC: fold change. FDR: false discovery rate. FC and FDR were calculated using the result of the alignment-based method, and labeled genera were supported by both pipelines. (**B**) Abundance of *Alphatorquevirus* and *Othohepavirus* in the alignment-based pipeline across different samples ranked in descending order; colors indicate different sample groups. (**C**) Virus genera with significant abundance alterations (FDR <0.05 and $\log_2$fold-change >1) in liver cancer patients when compared to HDs.

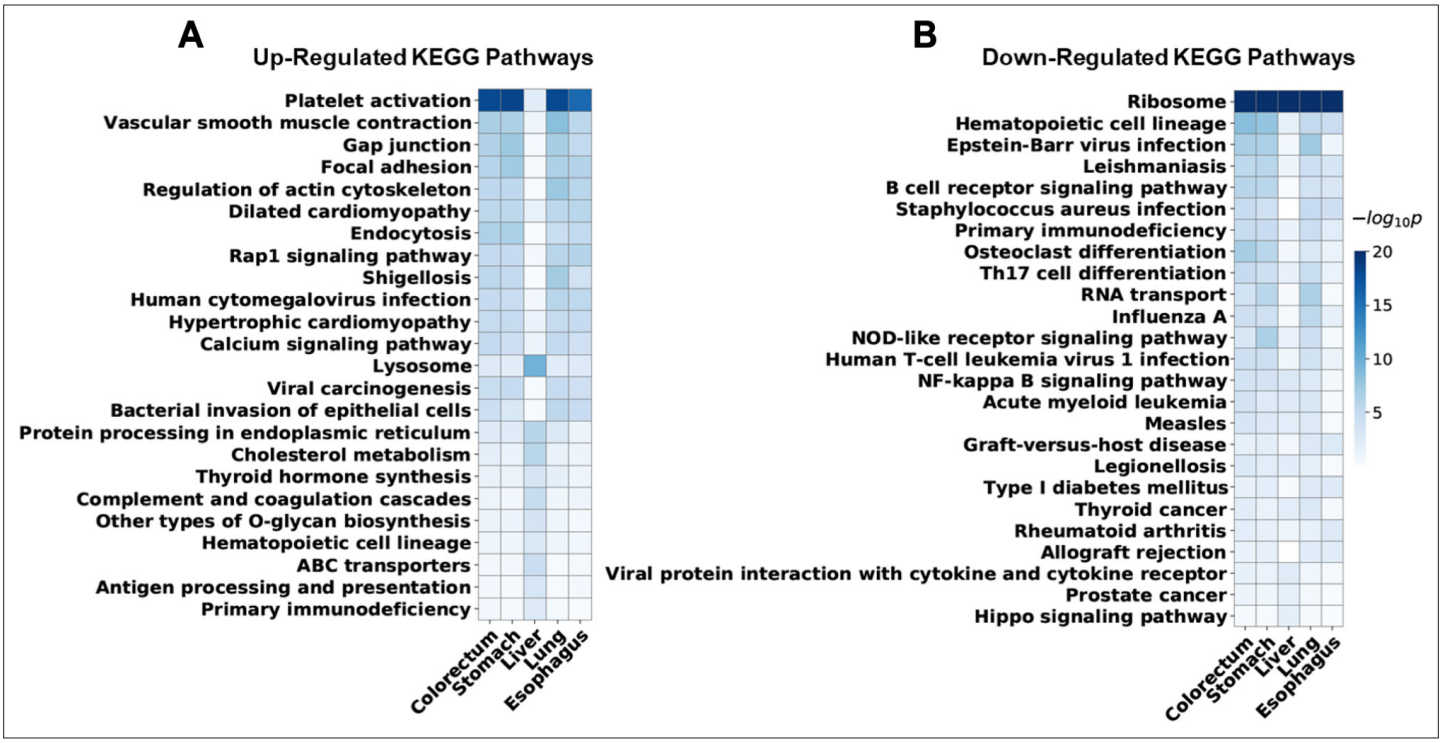

Figure 3—figure supplement 1. Enriched Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways of differentially expressed human genes for each cancer type. For five cancer types, the top 10 most significantly enriched KEGG pathways of upregulated and downregulated genes were identified. Union of these pathways was visualized. Rows for enriched pathways, columns for different cancer types, and colors indicate significance of the enrichment. Enriched pathways of liver cancer are relatively distinct from other cancer types, which may reflect some of its unique properties.

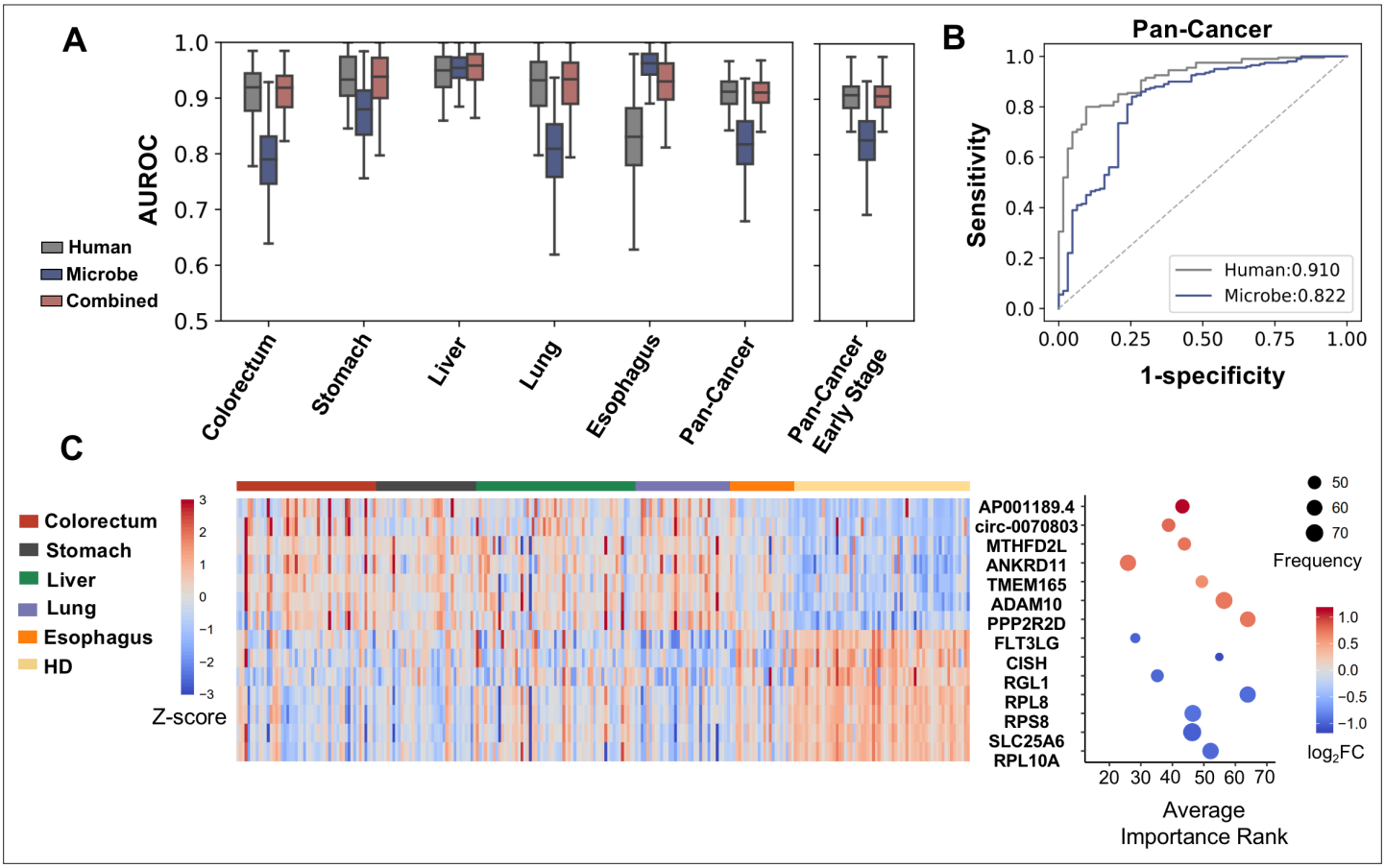

Figure 4. Cell-free RNA (cfRNA) features for cancer detection. (**A**) Performance (AUROC) on the holdout dataset in 100 rounds of bootstrap resampling using abundance of human gene expression, microbe abundance (kraken2's results), and combining both data for the binary classification (cancer patients vs. healthy donors). (**B**) Out-of-bag ROC curve using human or microbe features. For each sample, the median value of probabilities predicted by classifiers fitted in bootstrap replicates that reserved this sample in the testing set was utilized to generate the ROC curve. (**C**) Recurrent features with top fold changes when combining human and microbe features for bootstrap analysis. The left panel depicts Z scores of the expression levels in different subjects. The right panel illustrates their average importance ranks, frequency of identified as top 50 features, and fold change compared to healthy donors.

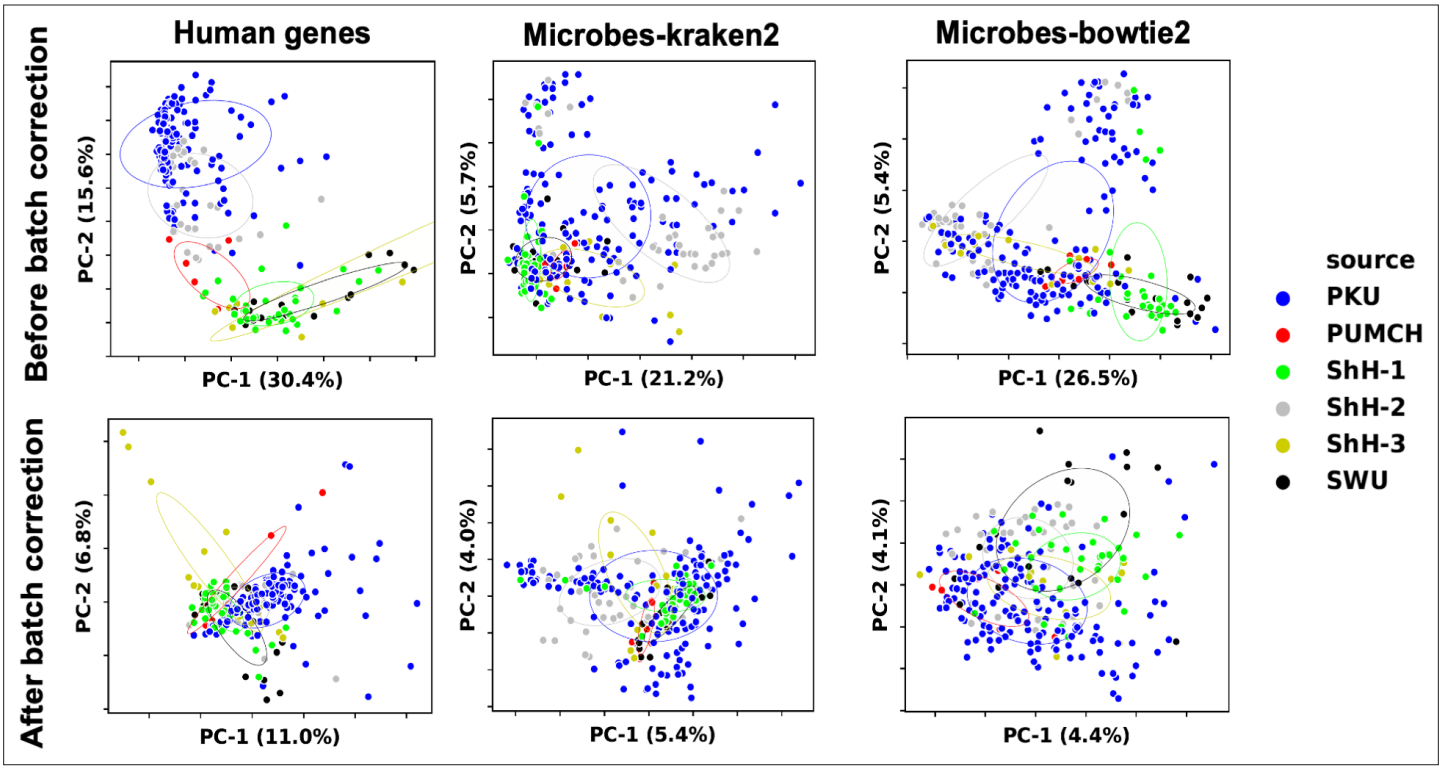

Figure 4—figure supplement 1. Data normalization for machine learning. We used RUVg to remove unwanted variations from trimmed mean of M-values (TMM) normalized gene expression and genus abundance. The 25% most insignificant features between different sample groups in discovery set reported by edgeR's ANOVA test were used as empirical controls. Data variations among different samples before (upper) and after (lower) RUVg processing were visualized with principal component analysis (PCA). Colors indicate different clinical centers.

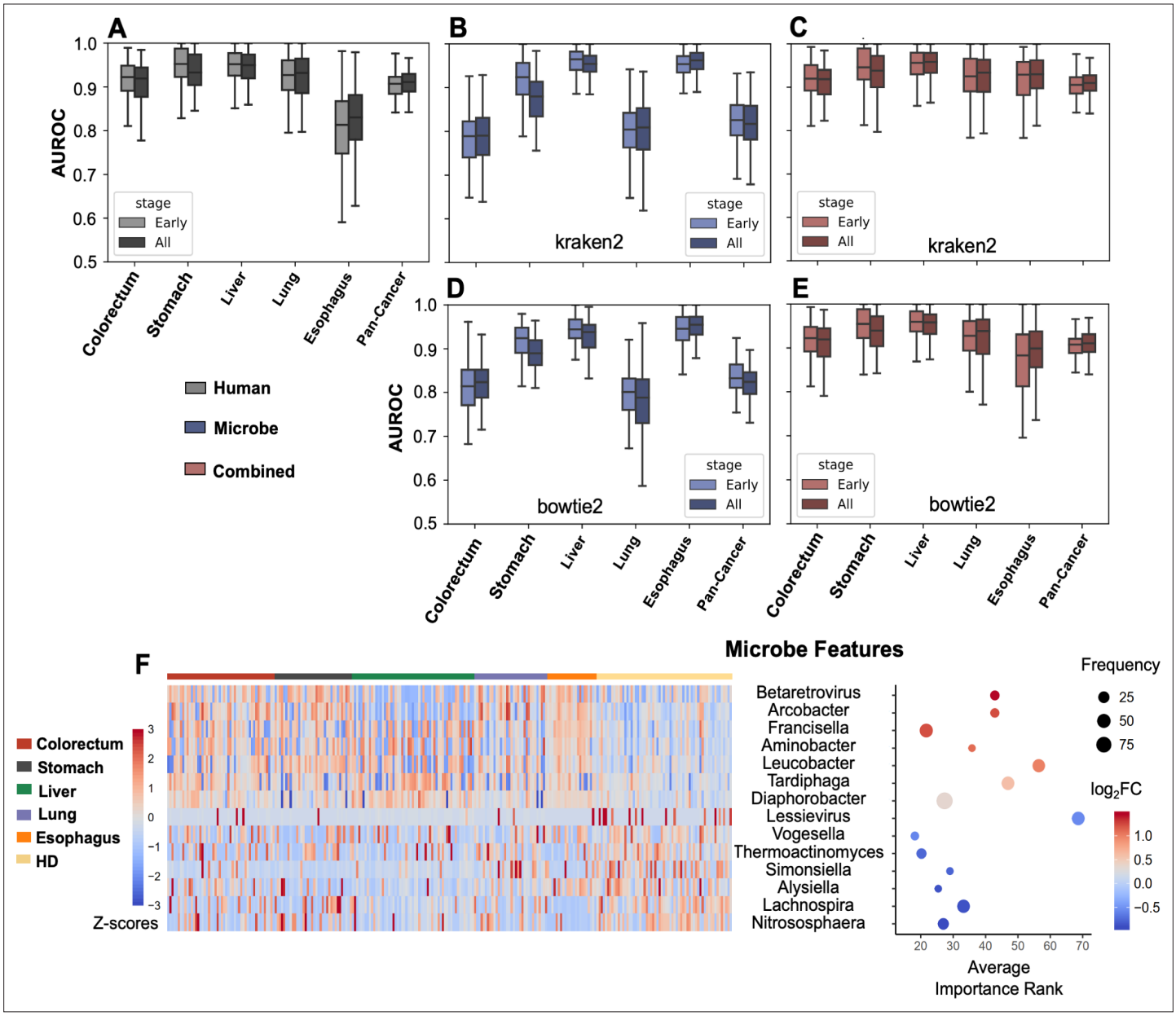

Figure 4—figure supplement 2. Binary classification for cancer detection. (**A–E**) Bootstrapping AUROC on holdout set. (**A**) Performance of human genes, stratified by cancer stages. (**B–C**) Performance of microbe features (**B**) and combined both microbe and human features (**C**) using kraken2's results. (**D–E**) Performance of microbe features (**D**) and combined both microbe and human features (**E**) using bowtie2's results. (**F**) Recurrently selected features with top fold changes when only considering microbe data for cancer vs. healthy donor (HD) classification.

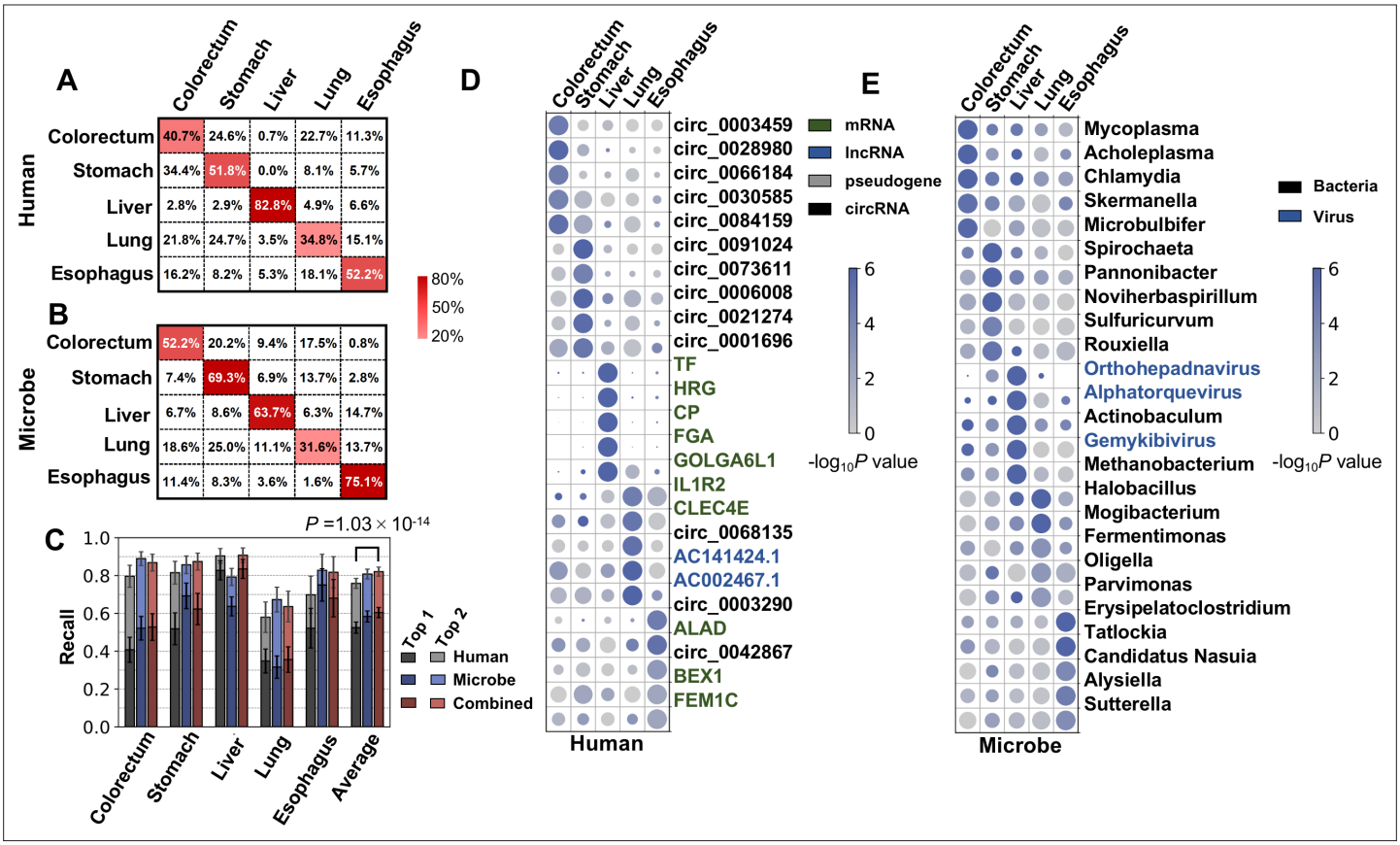

Figure 5. Cancer classification using human and microbial cell-free RNAs (cfRNAs). (**A–B**) Confusion matrix of human (**A**) and microbe (**B**) features averaged across bootstrap replicates. (**C**) Top 1 and top 2 recall for each cancer type in multiclass classification. The statistical significance was determined by a one-tailed Mann-Whitney U test. (**D–E**) Recurrent human (**D**) and microbe (**E**) features with the top fold change in multiclass classification. The sizes and colors of the circles indicate the relative abundances (bowtie2 result, scaled to 0–1) and p values in the one vs. rest comparisons, respectively.

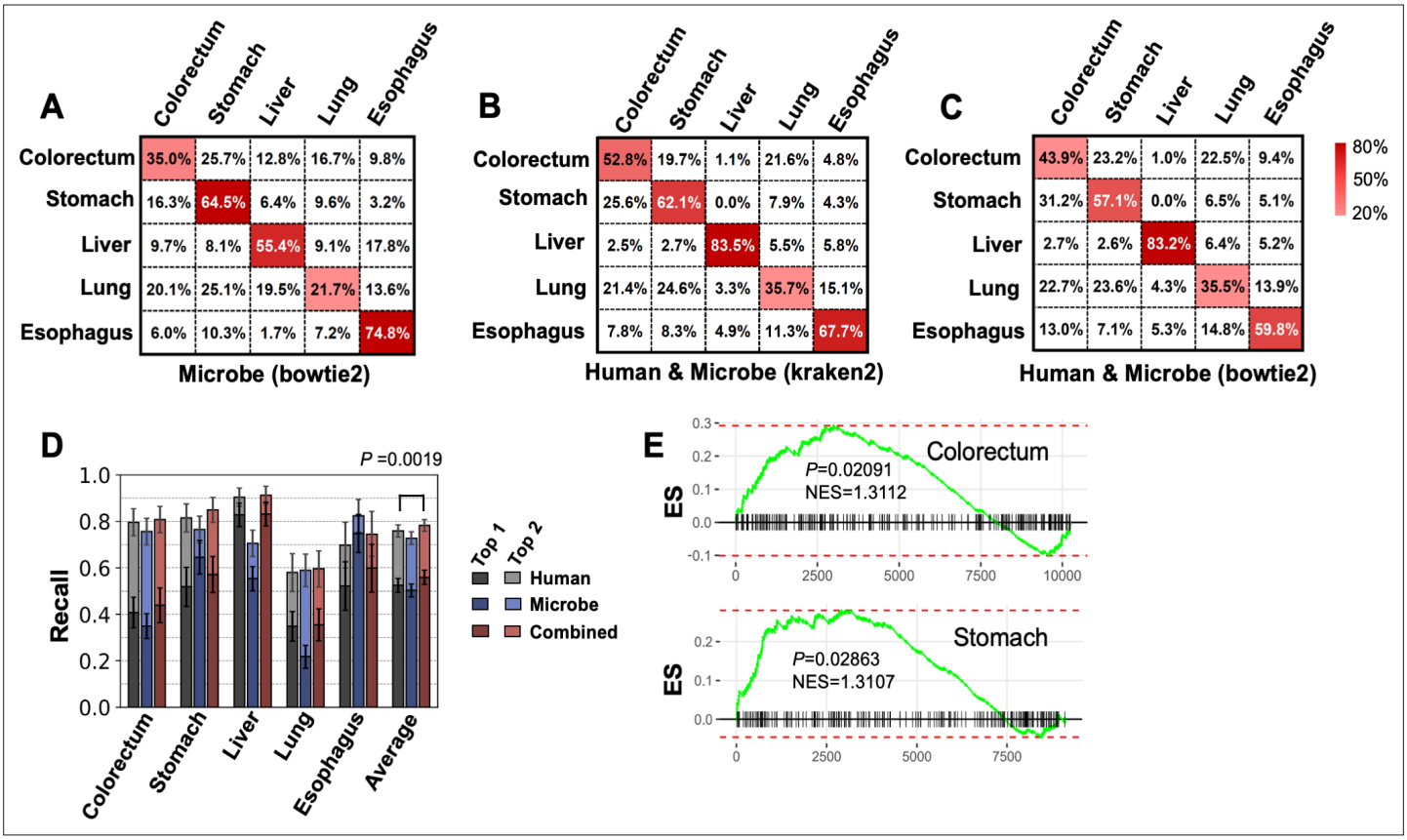

Figure 5—figure supplement 1. Performance for multiclass classification. (**A–C**) Confusion matrices for microbe features using bowtie2 pipeline (**A**), combine microbe and human features using kraken2 pipeline (**B**) and bowtie2 pipeline (**C**), and averaged across 100 bootstrap replicates. (**D**) Top 1 and top 2 recall of human and microbe features using bowtie2's results. The statistical significance was determined by one-tailed Mann–Whitney U test. (**E**) Gene set enrichment analysis (GSEA) for the enrichment of up to 300 circular RNAs (circRNAs) that upregulated in stomach cancer and colorectum cancer. circRNAs were ranked by fold change in tumor tissue vs. normal tissue comparison using mioncocirc data. ES: enrichment score; NES: normalized enrichment score.