CSI-Microbes: Identifying cell-type specific intracellular microbes from single-cell RNA-seq data

Welles Robinson[1,2,3*], Fiorella Schischlik[1], E. Michael Gertz[1], Joo Sang Lee[4], Kaiyuan Zhu[1,5], S. Cenk Sahinalp[1], Rob Patro[2,3], Mark D.M. Leiserson[2,3], Alejandro A. Schäffer[1^], Eytan Ruppin[1^*]

[1] Cancer Data Science Laboratory, Center for Cancer Research, National Cancer Institute, Bethesda, MD, USA 20892

[2] Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD, USA 20910

[3] Department of Computer Science, University of Maryland, College Park, MD, USA 20910

[4] Department of Artificial Intelligence & Department of Precision Medicine, School of Medicine, Sungkyunkwan University, Suwon 16419, Republic of Korea

[5] Department of Computer Science, Indiana University, Bloomington, IN, USA 47408

^ Equal contributing last authors

* Corresponding authors

Address for correspondence:

Dr. Eytan Ruppin

National Cancer Institute, National Institutes of Health

Building 15-C1

Bethesda, MD 20892

Eytan.Ruppin@nih.gov

Abstract

Despite recent reports of microbes living within tumor cells, the identification of intracellular microbes remains an open and important challenge. Here we introduce CSI-Microbes (computational identification of **C**ell type **S**pecific **I**ntracellular **Microbes**), a computational approach for the discovery of cell-type specific intracellular microbial taxa from single-cell RNA sequencing of host cells. It is first validated on two gold-standard datasets of human immune cells exposed to *Salmonella*. Next, CSI-Microbes is tested on Merkel cell and colorectal carcinoma where it identifies both reported and previously unknown tumor-specific intracellular microbes. Finally, CSI-Microbes is applied to analyze the intracellular microbiome of thirteen lung tumors where it identifies four tumors with bacterial taxa enriched in tumor cells and two tumors with bacterial taxa enriched in stroma or immune cells. Notably, the infected tumor cells down-regulate pathways associated with anti-microbial response and antigen processing and presentation, testifying to the functional significance of bacterial presence.

Introduction

Several recent papers have pointed to the functional importance of the tumor microbiome. For example, bacteria of the genus *Fusobacterium* are enriched in colorectal carcinoma compared to matched normal tissue, drive tumorigenesis, influence response to chemotherapy and bind to multiple human immune inhibitory receptors[1–6]. *pks+ E. coli* have been shown to induce a mutation signature frequently found in colorectal carcinoma[7]. In pancreatic cancer, a subset of taxa from the class *Gammaproteobacteria* were shown to mediate tumor resistance to chemotherapy[8]. A computational analysis of the unmapped reads from whole-genome sequencing (WGS) and whole-transcriptome sequencing (RNA-seq) experiments across 33 tumor types from The Cancer Genome Atlas (TCGA) cohort identified a variety of bacterial genera present in different tumor types and demonstrated that after filtering out potentially contaminant species, one can successfully build a predictor of cancer type based on tumors' microbial composition[9].

Some residents of the tumor microbiome have been shown to reside intracellularly within tumor and non-tumor cells in the tumor microenvironment[1,10]. The previously mentioned *Fusobacterium* has been shown to bind to ligands overexpressed at the surface of colorectal carcinoma cells, which it uses to invade these cells[1,11]. Another recent publication used multiple experimental techniques to interrogate the microbiome of seven cancer types[10]. It found that each cancer type has its own characteristic tumor microbiome and that many intratumoral bacteria exist intracellularly in both tumor and immune cells[10]. Further, it was recently reported that peptides derived from proteins in 41 bacterial species, including *Fusobacterium nucleatum*, are presented on the human leukocyte antigen class I and II (HLA-I and HLA-II) molecules of melanoma cells, which suggests that intracellular bacteria antigens can possibly be exploited

therapeutically[12]. Despite these advances, it has remained an important open challenge to identify which microbial taxa reside intracellularly and whether they reside exclusively or preferentially in tumor cells, immune cells, or cells of the non-cancerous solid tissue adjacent to the solid tumor (e.g., the study that identified the intracellular localization of some bacteria using staining also characterized the composition of the tumor microbiomes using 16S ribosomal RNA[13,14] but did not attempt to classify which bacterial taxa resided intracellularly in which cell types).

Here, we present a computational approach named **CSI-Microbes** (computational identification of **C**ell-type **S**pecific **I**ntracellular **Microbes**) that analyzes single cell RNA sequencing (scRNA-seq) datasets to identify intracellular microbes that are *cell-type specific* in a given tumor sample. Previous studies looking at microbial reads from scRNA-seq of host cells have focused on viruses[15,16]. To the best of our knowledge, the only previous study to analyze bacterial reads from scRNA-seq of host cells did so in the context of known *Salmonella* infection using a protocol designed to capture bacterial reads[17]. CSI-Microbes extends upon these studies by demonstrating that viruses and intracellular bacteria that preferentially reside within one cell-type can be discovered from two commonly used scRNA-seq protocols (Smart-seq2 and 10x) without knowing *a priori* the identity of the infecting virus or bacteria. It can do so if the input list of microbial genomes contains at least one genome with sufficient sequence similarity such that the microbe-derived reads can be aligned. In this discovery context, it is necessary to consider all microbial reads identified in the datasets, many of which are likely contaminants. Using user-specified cell-type annotations (such as those based on host transcriptomic data), CSI-Microbes identifies microbial taxa whose reads are enriched in specific cell types. Notably, this step controls for contaminating and extracellular microbes, whose

abundances is assumed not to vary significantly between cells of different types after proper normalization.

We demonstrate that CSI-Microbes correctly identifies known intracellular microbes and their cell-type preference in datasets of immune cells exposed to the intracellular bacterium *Salmonella enterica* and cancer cells with known or previously reported tumor cell-specific microbes. Next, we apply it to the unexplored intracellular microbiome of non-small cell lung carcinomas where we identified tumor cell-specific intracellular bacteria in four of the thirteen tumors. By comparing infected and uninfected tumor cells, we identify a transcriptomic signature of immune down-regulation in infected tumors cells, which both supports our findings of intracellular bacteria and suggests potential mechanisms for how and why intracellular bacteria reside within tumor cells.

Results

Overview of CSI-Microbes

The inputs to CSI-Microbes are (i) a database of microbial genomes and their taxonomy tree, (ii) FASTQ files from scRNA-seq experiments, (iii) cell metadata, including cell type annotations and (iv) known covariates, such as the sequencing plate, that may be associated with differential contamination. CSI-Microbes performs two tests for the identification of cell-type specific intracellular microbes: (a) *differential abundance*, which compares the normalized read counts of microbial taxa between cell types, and (b) *differential presence*, which compares the number of cells with at least one read from microbial taxa between cell types. We use the differential presence test specifically for the analysis of sparse 10x scRNA-seq datasets, which have few microbial reads, and the differential abundance test, otherwise. The output for each patient

sample is a list of candidate cell type-specific intracellular microbial taxa ranked by the statistical significance of their differential abundance or presence.

The algorithm proceeds in the following steps (**Fig. 1**, Methods): **(1) Identification:** scRNA-seq reads are mapped non-uniquely to microbial genomes using GATK PathSeq[18] after filtering reads aligned to the host genome and spike-in transcripts (differential abundance test only). **(2) Analysis:** For the differential abundance test, microbial reads are normalized across cells using spike-in sequences, log-transformed and compared across specified cell types using a two-sided Wilcoxon rank-sum test with minimum average $\log_2$ fold-change=.5. The statistical significance and the area under the receiver operating curve (AUC), which is equivalent to the U statistic of the Wilcoxon rank-sum test[19], are used to quantify the ability of the microbial taxa to discriminate between cell types, for each microbial taxon. For the differential presence test, the number of cells with at least one UMI from the microbial taxa are compared across specific cell types using a two-sided Fisher's exact test and the statistical significance and effect size (odds ratio) are reported[20]. Both tests can be applied to one or more taxa at the same level or along an ancestor-descendant path in the input taxonomy tree and corrected for multiple hypotheses. Both tests are run separately for cells given their covariate annotations (plate for Smart-seq2 and sample for 10x) and combined using Stouffer's Z-score method[21]. **(3) Validation:** Post-hoc validation tests of contamination inspired by decontam[22] are performed using spike-in reads and empty wells if available (Methods). These include two tests, the spike-in test and the empty wells test. The *spike-in test*, which is based on the observation that the number of reads from contaminating microbes are likely to correlate inversely with the sample DNA concentration, calculates the correlation between the spike-in reads and the reads of the taxon of interest. The *empty wells test*, which is based on the observation that contaminating sequences are more likely

to show up in negative controls, compares the presence of microbial taxa between empty and non-empty wells.

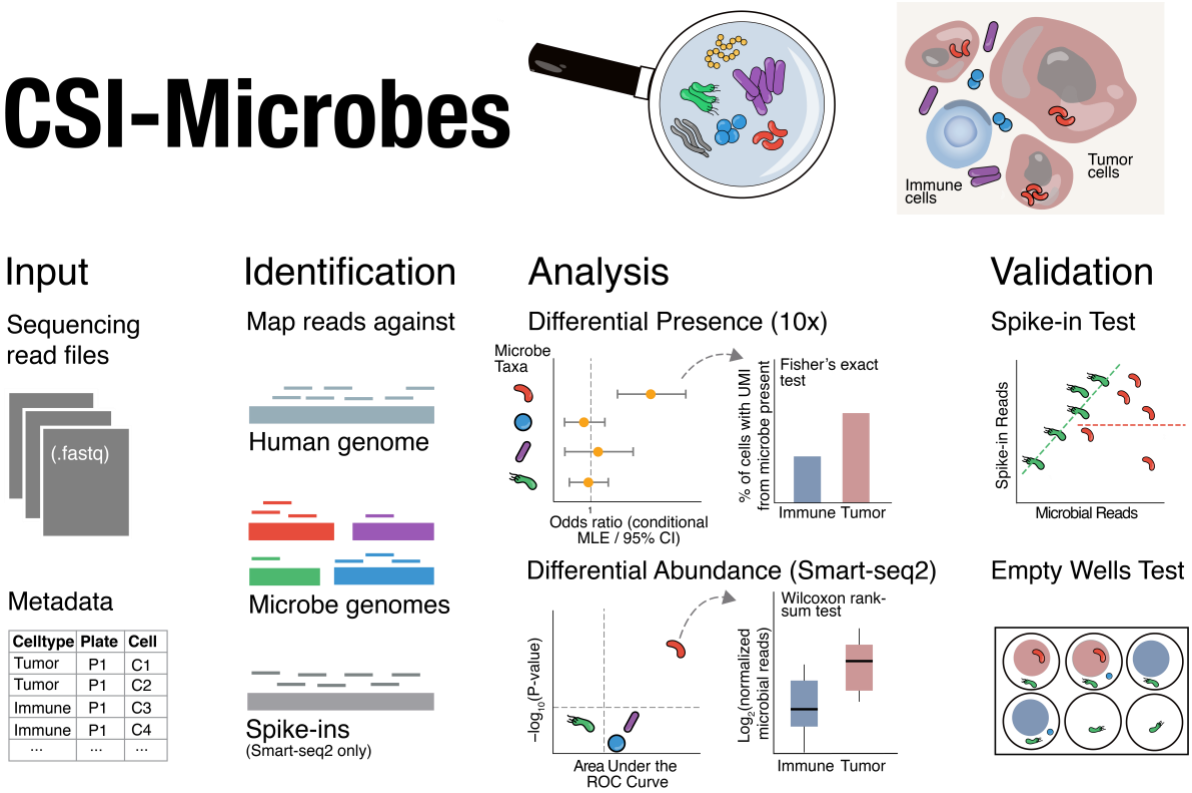


Figure 1: *Overview of the CSI-Microbes approach, which takes sequencing read files (fastq format) from scRNA-seq experiments (10x or Smart-seq2) and cell metadata such as cell-type and plate (Smart-seq2) or sample (10x) as input. The steps are (1) Identification: reads are mapped to the human genome and optionally spike-in sequences (Smart-seq2 only) and unmapped reads are mapped against a large set of microbial genomes; (2) Analysis: for microbial taxa with reads identified in step 1, the $\log_2$ spike-in normalized reads (Differential Abundance) or the number of cells with at least one UMI (Differential Presence) are compared between user-specified cell-types (provided as input) while controlling for covariates associated with contamination (also provided as input); (3) Validation: for microbial taxa, the correlation between the number of reads from this taxa and the reads from spike-in sequences is reported (Spike-in Test) and the frequency of reads from this taxa is compared between empty wells and wells with cells (Empty Wells Test).*

Applying CSI-Microbes to the *Salmonella* gold-standard validation scRNA-seq datasets

We validated the differential abundance test of CSI-Microbes on a "gold-standard" Smart-seq2 dataset that sequenced 262 human monocyte-derived dendritic cells (moDCs) that were exposed to either the *D23580* strain or the *LT2* strain of *Salmonella enterica* as well as 80 control cells[23]. The 262 *Salmonella* exposed cells were further labeled as 135 "infected" and 127 "bystander" cells depending on whether the presence of live, intracellular *Salmonella* could be detected in them using FACS. The identification step (step 1) found a median of 8,030 bacterial reads per cell that mapped to 859 bacterial genera including *Salmonella*. We applied the spike-in test (step 3) to the 19 most abundant genera and found that, indeed, the number of reads from the 18 genera other than the expected *Salmonella* is positively correlated with the number of spike-in reads, suggesting that these 18 other genera are contaminants. CSI-Microbes identified only the different microbial taxa along the path from the class *Gammaproteobacteria* (p-value=9e-6, AUC=.66) to the species *Salmonella enterica* (p-value=1e-8, AUC=.70) as differentially abundant between the infected and bystander cells (**Fig. 2A**). We observed false positives when comparing cells across plates, illustrating the importance of controlling for the sequencing plate (step 2 in CSI-Microbes, Supplementary Information).

We next validated the differential presence test of CSI-Microbes on a 10x dataset in which the authors sequenced 3,485 human peripheral blood mononuclear cells (PBMCs) that were exposed to *Salmonella enterica* serovar Typhimurium strain *SL1344*[24]. Using flow cytometry, the authors determined that ~3% of the exposed PBMCs, including 90% of the monocytes, were infected with live red fluorescent protein (RFP)-expressing intracellular *Salmonella*. We applied CSI-Microbes to look for differentially present microbes between the monocytes and non-monocytes, which identified the path from the phylum *Proteobacteria* to the

genus *Salmonella*. This was achieved even though only 29 UMIs mapped to bacterial genomes in this dataset (**Fig. 2B**).

We validated our findings of differential abundance and differential presence in the respective *Salmonella* datasets using two alternative mapping approaches (step 1). First, we reproduced our findings at the genera level in both datasets using the *k*-mer tool CAMMiQ[25] instead of the alignment-based approach PathSeq. Second, we directly mapped unaligned reads to the *Salmonella* strain genome used in each experiment using the error-tolerant and ambiguity-character-tolerant aligner SRPRISM[26] (Supplementary Information and **Supplementary Fig. 1 and 2**). The SRPRISM analysis of the Smart-seq2 dataset testifies to the importance of aligning reads to the entire microbial genome, instead of specific regions like the commonly used 16S rRNA loci, which contain only ~9% of the total reads mapped to the *Salmonella* strain genomes (Supplementary Information).

In these two gold-standard datasets, CSI-Microbes identified enrichments at different levels of the NCBI Taxonomy tree. In the Smart-seq2 dataset, it identified microbial taxa from the class *Gammaproteobacteria* to the species *Salmonella enterica* while in the 10x dataset, it found the path from the phylum *Proteobacteria* to the genus *Salmonella* (but not the species *Salmonella enterica*). To programmatically report the highest resolution microbial taxa, CSI-Microbes analyzes microbial reads from the class to the species level while performing hierarchical false discovery rate (hFDR) correction to limit the number of hypotheses[27,28].

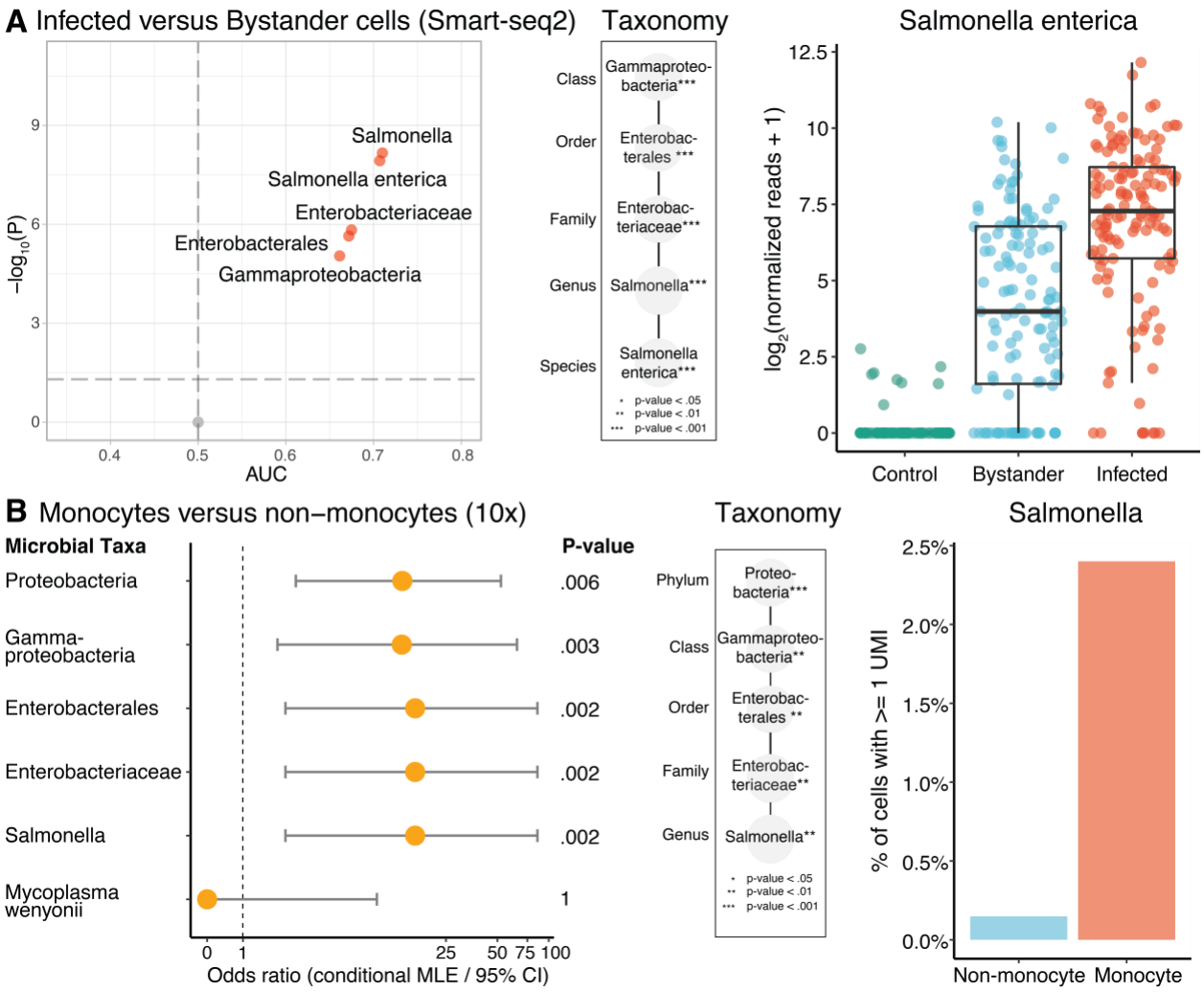


Figure 2: ***(A)*** *Overview of the results from CSI-Microbes on the Smart-seq2 dataset [23]. The volcano plot shows the microbial taxa found to be more differentially abundant (p-value < .05 and AUC > .5) in infected cells compared to bystander cells (**Supplementary Table 1**). The taxonomy illustration shows that all differentially abundance microbial taxa (shown on the left) are either Salmonella enterica or its taxonomic ancestors. The box plot shows that there are significantly more normalized reads mapped to Salmonella enterica in infected cells compared to bystander cells, which have significantly more normalized reads than control cells, which were not exposed to Salmonella enterica and contain almost no reads mapped to Salmonella enterica.* ***(B)*** *Overview of the results from CSI-Microbes on sample GSM3454529 (exposed to Salmonella enterica) from the 10x dataset[24]. The odds ratio plot shows that Salmonella and its taxonomic ancestors (see taxonomy illustration) but not Mycoplasma wenyonii are more differentially present in monocytes compared to non-monocytes (p-value < .05 and odds ratio > 1) (the taxonomic ancestors of Mycoplasma wenyonii, which received identical scores, were excluded for space purposes, and are listed in **Supplementary Table 1**). The third panel shows that UMIs from Salmonella were identified more frequently in monocytes compared to non-monocytes.*

Application of CSI-Microbes to Merkel cell and colon carcinomas

We next validated CSI-Microbes in the context of cancer by analyzing two 10x scRNA-seq datasets from two tumor types previously reported to have tumor cell-specific intracellular microbes. Previous reports have estimated that ~80% of Merkel cell carcinomas are driven by the clonal integration of the Merkel polyomavirus[29]. We applied CSI-Microbes to identify differentially present microbes between tumor and non-tumor cells from two Merkel cell tumors[30], which were confirmed to have integration of the Merkel polyomavirus by the original authors. In both patients, CSI-Microbes identifies the species *Human polyomavirus* 5, for which the only fully sequenced genome comes from the "no rank" child taxon *Merkel polyomavirus*, to be differentially present in tumor cells (patient 2586-4 pre-treatment: p-value=7×10^{-9}, odds ratio=5.3, post-treatment: p-value=.04, odds ratio=6; patient 9245-3: p-value=10×10^{-64}, odds ratio=12.8) (**Supplementary Table 2** and **Supplementary Fig. 3**).

The bacterial species *Fusobacterium nucleatum* has previously been reported to preferentially reside within colorectal carcinoma cells and to a lesser extent, stromal cells[1]. We applied the differential presence test of CSI-Microbes to compare tumor and non-tumor cells from a recently published 10x dataset of six colorectal carcinomas, which were not evaluated for the presence of bacteria by the authors[31]. In the tumor sample from patient SC028 (the only sample with UMIs aligned to *Fusobacterium*), CSI-Microbes identified the genus *Fusobacterium* (p-value=.028, odds ratio=2.7) to be more differentially present in the tumor cells compared to non-tumor cells in patient SC028 (**Fig. 3A**). We observed that the two cells with the most UMIs from *Fusobacterium* (24 and 16 UMIs respectively) were stromal cells, suggesting the presence of intracellular *Fusobacterium* in a small percentage of stromal cells as has been previously suggested[1]. CSI-Microbes also identified the bacterial species *Hathewaya histolytica* (p-

value=.01, odds ratio=3.6) (previously called *Clostridium histolyticum*) to be differentially present in the tumor cells from patient SC019 (**Fig. 3B** and **Supplementary Table 3**). Interestingly, *Hathewaya histolytica* has been reported to be strongly enriched in the colonic tissue of patients with ulcerative colitis compared to controls[32].

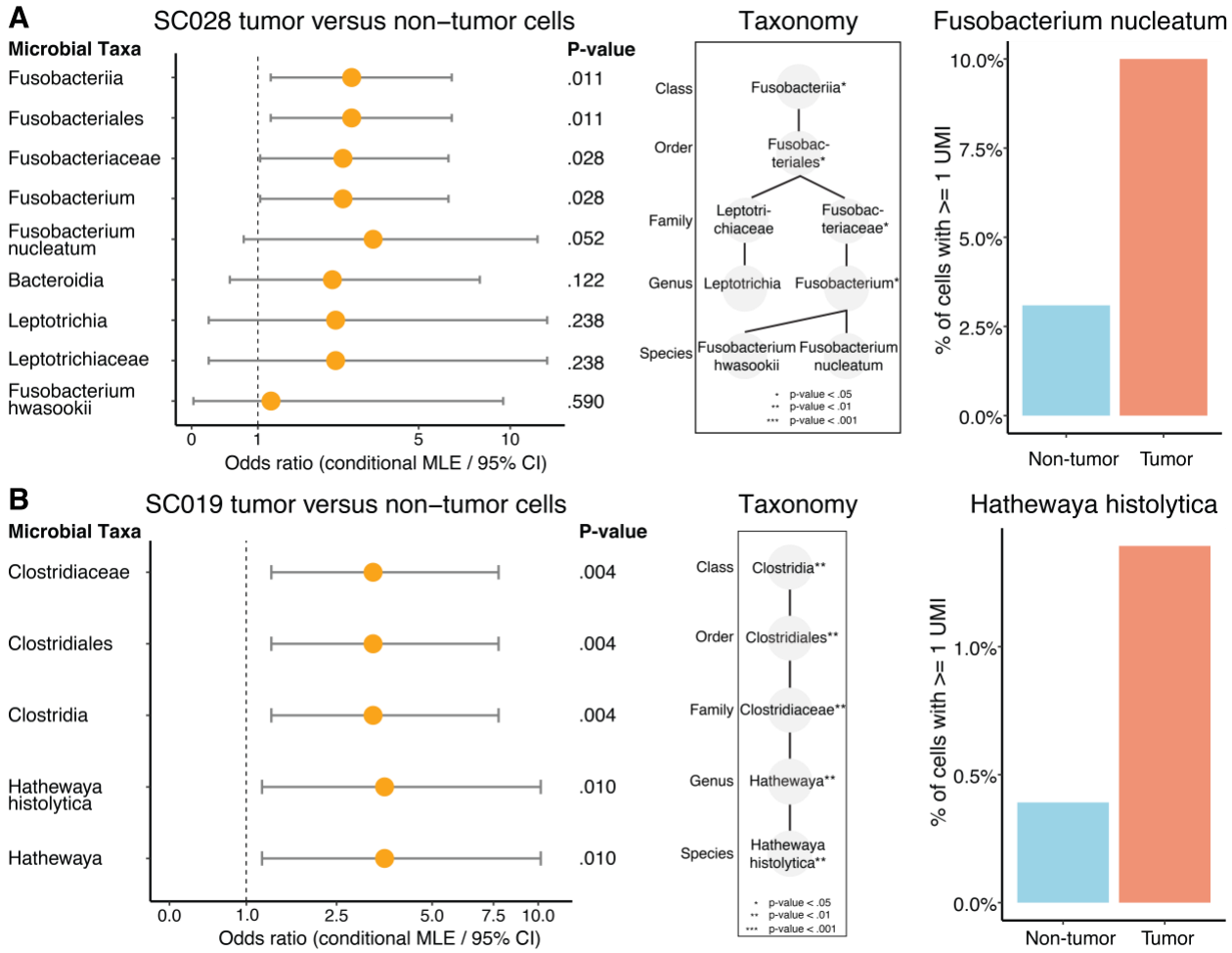


Figure 3: *Overview of significant results from CSI-Microbes on colorectal carcinoma 10x dataset[31]. **(A)** The odds ratio plot shows Fusobacterium and its taxonomic ancestors to be more differentially present in tumor cells compared to non-tumors in the tumor sample from SC028 (see **Supplementary Table 3** for more details). Although not statistically significantly, the barplot shows the difference between the percentage of tumor and non-tumor cells with at least one UMI from Fusobacterium nucleatum. **(B)** The odds ratio plots shows that Hathewaya histolytica and its taxonomic ancestor are more differentially present in tumor cells compared to non-tumor cells from SC019. UMIs from Hathewaya histolytica are found in a higher percentage of tumor cells compared to non-tumor cells.*

Application of CSI-Microbes to lung cancer

Next, we applied CSI-Microbes to identify differentially abundant microbes from a large, recently published lung cancer Smart-seq2 scRNA-seq dataset with spike-in sequences[33]. We analyzed 13 lung cancer tumors where at least 10 tumor cells and 10 non-tumor cells were sequenced in the same plate, comprising in total ~11,000 cells from 50 sequencing plates. We first compared the microbial reads (from the identification step of CSI-Microbes) to a previous analysis of the lung microbiome using 16S rRNA[10]. We identified at least one unambiguous read to 16 of the 17 species with complete genomes that were reported to be enriched in lung cancer, suggesting that Smart-seq2 data may provide sufficient coverage of the tumor microbiome (Methods). Next, using the authors' cell-type annotations (tumor, immune, stroma and epithelial), we applied CSI-Microbes to identify multiple tumors where microbial taxa are differentially abundant in tumor cells compared to immune cells (TH231, TH236, TH238, TH266; see **Fig. 4A,B** for examples and **Supplementary Table 4** for the complete list) and stromal cells (TH236, TH266). We also detected two tumors with taxa that are differentially abundant in stromal cells (TH231; **Fig. 4C**) or immune cells (TH220) compared to the tumor cells. Notably, all four tumor samples containing tumor cells enriched with bacterial taxa are from tumors that had undergone at most one prior drug treatment. In contrast, the tumor sample with bacterial taxa enriched in immune cells came from a patient who had six prior lines of treatment including immunotherapy.

CSI-Microbes identified the species *Cutibacterium acnes* to be differentially abundant in the tumor cells compared to the immune cells in four of the thirteen tumors (TH231, TH236, TH238, TH266). *Cutibacterium acnes* was excluded from a previous experimental exploration of the lung tumor microbiome[10] because it was identified in a large percentage of the tissue-free

negative controls, which indicated that it is a contaminant. Consistent with this finding, we identify reads from *C. acnes* in nearly every single cell analyzed. However, *C. acnes* is significantly more abundant in tumor cells compared to immune cells in all four tumors (and is not significantly more abundant in non-tumor cells in any other tumor). Notably, *C. acnes* has been previously reported to exist intracellularly in epithelial cells and to be a commensal in the lung and skin as well as an opportunistic pathogen[34–36]. Thus, unlike bulk sequencing-based computational methods, CSI-Microbes can consider all microbes while implicitly controlling for contaminants by comparing normalized microbial reads between cells of the same patient. CSI-Microbes identifies another member of the *Cutibacterium* genus, *Cutibacterium granulosum*, which was reported to be enriched in lung cancer[10], as differentially abundant in tumor cells in patient TH266 (uncorrected p-value = .04, **Fig. 4B**). Additional taxa that are differentially abundant in tumor cells include the genera *Corynebacterium* (TH236 and TH238) (**Fig. 4A**) and *Staphylococcus* (TH236) and the family *Micrococcaceae* (TH238) (**Fig. 4A**). In patient TH231, where CSI-Microbes found *C. acnes* to be enriched in tumor cells, it also identified the genus *Leptotrichia* to be enriched in stroma cells, which were further classified as melanocytes by the original authors, compared to non-stroma cells (**Fig. 4C**). In patient TH220, CSI-Microbes identified both the *Micrococcus* and *Corynebacterium* genera and the *Alphaproteobacteria* class to be enriched in immune cells compared to tumor cells in patient TH220. We did not find any bacterial taxa to be differentially abundant between macrophages/monocytes and any other cell type.

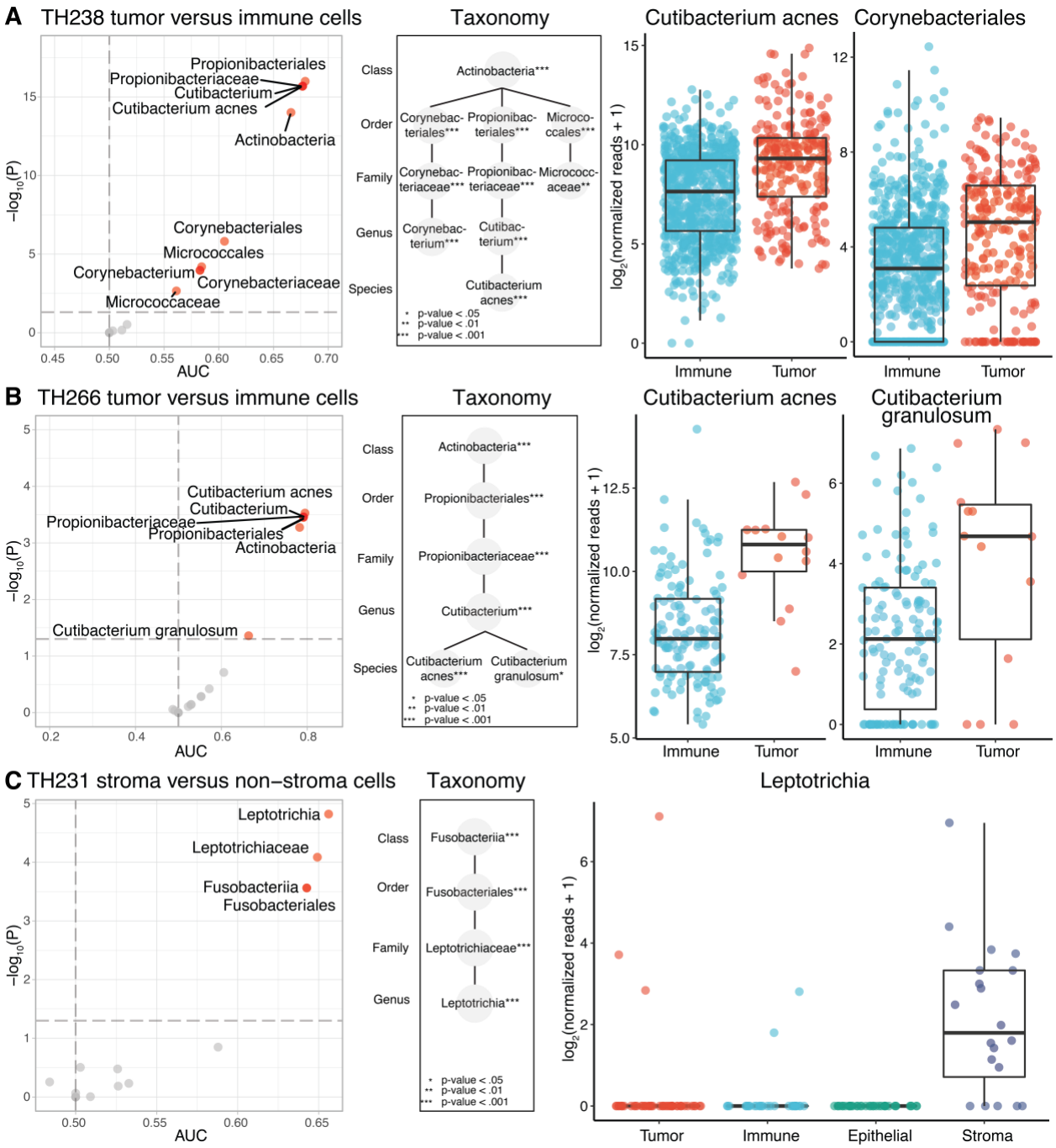


Figure 4: *Key results from CSI-Microbes analysis of lung cancer* ***(A)*** *CSI-Microbes identified multiple bacterial taxa (all descendants of the Actinobacteria class) to be enriched in tumor cells compared to immune cells in patient TH238 (AUC > .5 and p-value < .05).* ***(B)*** *CSI-Microbes identified two species from the Cutibacterium genus (C. acnes and C. granulosum) to be enriched in tumor cells compared to immune cells in patient TH266.* ***(C)*** *CSI-Microbes identified the bacterial genus Leptotrichia to enriched in stroma cells compared to non-stroma cells in patient TH231. The boxplot shows that there are more reads to Leptotrichia in stroma cells compared to epithelial, immune and tumor cells.*

To study the transcriptomic state associated with the presence of intracellular bacteria in the lung cancer cohort that we analyzed, we performed a differential expression analysis between the tumor cells from patients TH231, TH236, TH238 and TH266 (termed "infected" because CSI-Microbes identified microbial taxa that are differentially abundant in tumor cells in each of these samples) and the tumor cells from the other patients (termed "uninfected") (**Fig. 5A** and **Supplementary Table 5** and Methods). The most down-regulated gene in infected tumor cells compared to uninfected tumors cells is *S100A9* (FDR-corrected p-value=1e-62, AUC=.09, where for down-regulated genes 0 indicates perfect discrimination). *S100A9* binds to *S100A8* to form a heterodimer calprotectin, which has antimicrobial properties due to its ability to sequester metal ions such as zinc, manganese and iron that are essential nutrients for microbes[37]. The down-regulation of *S100A9* thus may explain how bacteria such as *C. acnes* can survive inside these tumor cells.

Next, we performed a gene set enrichment analysis (GSEA) of the differentially expressed genes between the infected and uninfected cancer cells and clustered similar gene sets using Enrichment Map[38] (**Fig. 5B** and **Supplementary Table 6** and Methods). The largest cluster of gene sets down-regulated in infected tumor cells contains predominately gene sets associated with processing and presentation of antigens as well as those associated with hematopoietic differentiation and response to external stimulus. This cluster is connected to the chemotaxis cluster, which includes gene sets associated with chemotaxis of leukocytes, granulocytes, and neutrophils. We observed at least three additional and unconnected down-regulated gene sets that we could associate with anti-microbial response, including *humoral immune response mediated by antimicrobial peptides*, *transition metal ion homeostasis* and *cell killing*. Additional immune response gene sets including *response to interferon gamma* and

response to interferon beta as well as *interleukin-12 production* are also down-regulated in the infected tumor cells. The largest cluster of up-regulated gene sets includes many gene sets associated with microtubules, which have previously been shown to be modulated by intracellular pathogens[39]. The association of intracellular bacteria with the down-regulation of the antigen presentation system in tumor cells, which has been previously observed in one *Salmonella* dataset[23] (see also Supplementary Information) is particularly relevant given the recent finding that peptides derived from bacteria can be present on the HLA class I and II molecules in melanoma[12].

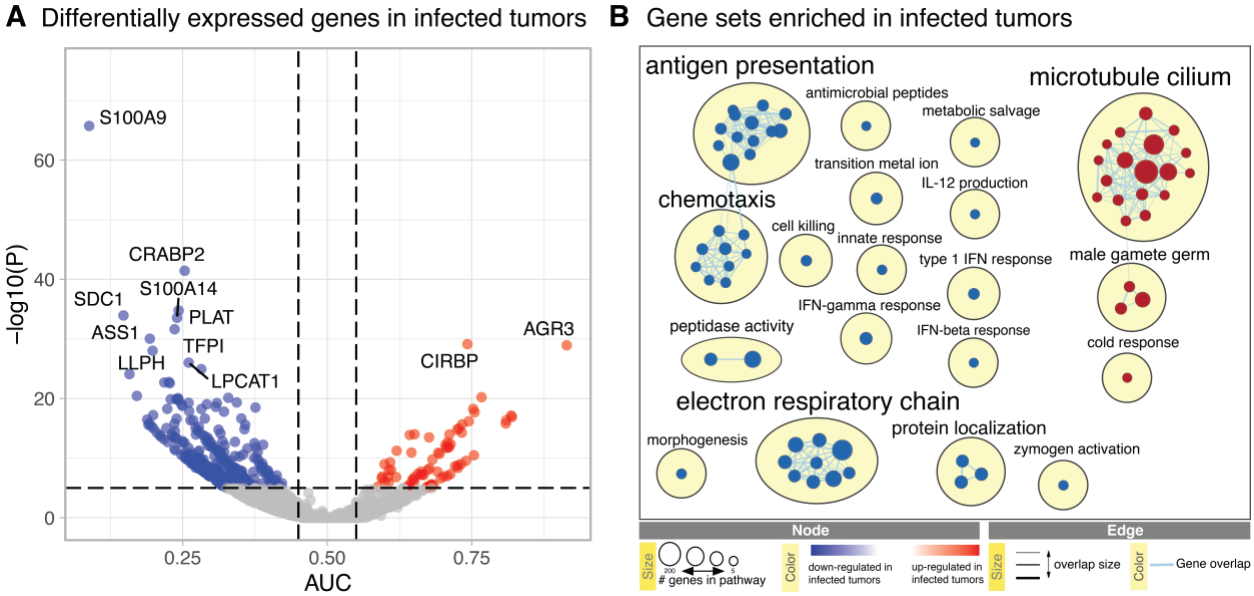


Figure 5: (A) *Volcano plot of the differentially expressed genes (**Supplementary Table 5**) between the infected and uninfected tumors cells where blue genes (AUC < .45 and p-value < 10e-6) are significantly down-regulated in infected tumor cells and red genes (AUC > .55 and p-value < 10e-6) are significantly up-regulated in infected tumor cells.* ***(B)*** *Enrichment map of the significantly enriched gene sets (FDR q-value < .02) (**Supplementary Table 6**) where blue nodes represent gene sets that are down-regulated in infected tumor cells and red nodes represent gene sets that are up-regulated in infected tumor cells. Gene sets are connected via edges if there are many genes in common between the two gene sets. Similar gene sets are clustered and manually named using common terms.*

Discussion

We developed and validated a first of its kind computational tool for the discovery of cell-type-specific intracellular microbes from single-cell transcriptomics data. Our approach can be applied to many existing scRNA-seq datasets without additional wet lab experiments because our framework of differential presence and abundance controls for contamination. Our bioinformatic approach is complementary to and extends existing wet-lab approaches for identifying intracellular microbes in the tumor microenvironment. Complementing HLA peptidomics, single cell transcriptomic analysis identifies not only likely intracellular bacteria but also their cell-type of residence. CSI-Microbes simultaneously searches all taxa in the input database while a staining-based approach like RNAscope[40], which can identify cells containing intracellular bacteria with high confidence, is limited to using probes that are either specific to a small, pre-determined set of taxa (requiring *a priori* knowledge) or generic to all bacteria like 16S (hiding the identity of the intracellular bacteria). Intracellular microbes reported by CSI-Microbes can provide the *a priori* knowledge needed by approaches like RNAscope, which can and should be used to experimentally validate scRNA-seq-based findings. We also expect our approach to improve as additional tumor resident microbes are fully sequenced because many species found to be present in the tumor microbiome using 16S rRNA sequencing[10] lack complete reference genomes (Methods).

Beyond our specific approach, we suggest that scRNA-seq is particularly well-suited for identifying intracellular bacteria for three reasons. First, in contrast to bulk sequencing approaches, scRNA-seq protocols provide many “replicates” (in the form of single cells) from the same individual often from multiple batches, which can be used to identify and control for contaminating microbes. Second, scRNA-seq protocols are designed to sequence the RNA inside

of single cells, which includes the RNA from any intracellular microbes (although many existing protocols, including those analyzed in this study, may under sample microbial RNA for various technical reasons). Third, scRNA-seq contains information on host transcriptomic changes associated with the presence of intracellular microbes. The findings of intracellular bacteria within tumor cells reported here and by others[10,12] raise questions concerning the putative functional roles of these intracellular microbes: are they simply "innocent bystanders" and opportunistic pathogens or do they play important functional roles in tumorigenesis and response to treatment? By analyzing both human and microbial reads, we found the down-regulation of antigen processing and presentation in both moDCs infected with intracellular *Salmonella* (Supplementary Information) as already suggested[23] and infected tumor cells in NSCLCs. These findings both provide additional support for our results and point to a potential win-win relationship between intracellular bacteria and the tumors that host them: intracellular bacteria down-regulate the antigen processing and presentation system of the host cell, which helps the tumor evade the immune system.

We focus on cancer-specific intracellular microbes, but we note that CSI-Microbes can be applied to identify intracellular microbes in other diseases. Although many intracellular bacteria preferentially invade one cell-type[41], many other intracellular microbes may either not be cell-type-specific or sufficiently enriched in one cell-type to be identified by our tests. We note that the combination of our spike-in test and empty wells test could be used to identify a more general class of intracellular microbes although we did not pursue this approach because none of the datasets analyzed contained more than a small number of empty wells. In such future applications, our results underscore the importance of using spike-in sequences, empty wells, and

multiple cell-types in the same plate to further enhance the detection accuracy of intracellular bacteria from sequencing data.

We expect that the performance of CSI-Microbes will improve as scRNA-seq technology advancements continue to increase the number of cells and sequencing depth and overcome some of the technical reasons, such as the exclusive use of polyA capture, that causes many existing protocols to under sample prokaryotic RNA molecules[17]. The use of polyA tail selection to enrich for polyadenylated eukaryotic mRNAs selects against prokaryotic RNA molecules, that have shorter polyA tails when polyadenylated and are less likely to be polyadenylated[42]. It was shown nearly thirty years ago that polyadenylation is carried out by the polymerase PAP1 in *E. coli*, which is a member of the *Enterobacteriaceae* family[43]. Although it is not known whether all bacterial taxa have polyadenylated RNAs, we hypothesize that any taxon with an orthologue of PAP1 will have some polyadenylated mRNA. It is also largely unknown which bacterial genes are polyadenylated as there are only a small number of *Enterobacteriaceae* genes have been shown to be polyadenylated in *E. coli*[44]. Using our results from aligning the unmapped Smart-seq2 reads against the genome of *Salmonella,* which is also a member of the family *Enterobacteriaceae,* we can add to the list of likely polyadenylated genes in this family **(Supplementary Table 1).**

In summary, CSI-Microbes is a powerful *in silico* approach for analyzing the intracellular tumor microbiome. We have shown that it identifies both known and previously reported cell-type-specific intracellular bacteria and viruses as well as multiple instances of tumor-specific intracellular bacteria in NSCLCs associated with a transcriptional signature of immune and anti-microbial response down-regulation.

Methods

Data Availability

In this study, we analyzed publicly available FASTQ files from the following datasets:

Sample Type	Patients analyzed (total)	Cells analyzed	Approach	Reference	Spike-ins	Empty Wells
Dendritic cells (exposed to *Salmonella*)	1 (1)	342	Smart-seq2	NCBI BioProject PRJNA437328[23]	Yes (ERCC)	2
PBMCs (exposed to *Salmonella*)	1 (1)	3,485	10x	NCBI BioProject PRJNA503437[24]	NA	NA
Merkel cell carcinoma	2 (2)	12,754	10x	NCBI BioProject PRJNA483959 (patient 2586-4), PRJNA484204 (patient 9245-3)[30]	NA	NA
Colorectal carcinoma	6 (29)	27,414	10x	ArrayExpress E-MTAB-8410[31]	NA	NA
Non-small cell lung carcinoma	13 (30)	10,562	Smart-seq2	NCBI BioProject PRJNA591860[33]	Yes (ERCC)	0

For technical reasons, 10x scRNA-seq experiments do not use spike-in sequences or empty wells. We excluded 23 of the 29 colorectal carcinomas[31] (the Korean cohort) from our analysis because the raw reads were not publicly available. We excluded 17 of the 30 non-small cell lung carcinomas (NSCLCs)[33] from our analysis because either these patients did not have cells sequenced using spike-in sequences (2 patients) or these patients did not have at least one sequencing plate with at least ten tumor cells and ten non-tumor cells (15 patients).

Code Availability

CSI-Microbes is logically partitioned into two modules, one module for the "identification" step and one module for the "analysis" and "validation" steps. A reproducible Snakemake[45] (https://snakemake.readthedocs.io/) workflow for identifying microbial reads from scRNA-seq datasets, which includes code to download the data from the datasets above, is available on GitHub (https://github.com/ruppinlab/CSI-Microbes-identification) although the identification module has some dependencies to the NIH Biowulf server. A reproducible Snakemake workflow for analyzing microbial reads to identify differentially abundant or differentially present microbes is available on GitHub (https://github.com/ruppinlab/CSI-Microbes-analysis). To facilitate reproduction of our analyses, we have uploaded CSI-Microbes-analysis v0.2.0 (the version used in this study) along with the relevant input files (generated by CSI-Microbes-identification v0.2.0) to Zenodo (https://doi.org/10.5281/zenodo.4695248).

Smart-seq2 datasets preprocessing

Raw FASTQ files were trimmed using fastp[46] v0.20.1 with the arguments "*--unqualified_percent_limit 40 --cut_tail --low_complexity_filter --trim_poly_x*". The trimmed FASTQ files were aligned to the reference human genome (GRCh38 gencode release 34) and any applicable spike-in sequences using STAR[47] 2.7.6a_patch_2020-11-16 with the arguments "--soloType SmartSeq --soloUMIdedup Exact --soloStrand Unstranded --outSAMunmapped Within".

10x datasets preprocessing

Raw FASTQ files were aligned to the reference human genome using CellRanger[48] v5.0.1 (https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger). The annotated polyA and template sequence oligonucleotide (TSO) sequences were trimmed, the unmapped reads were converted to the FASTQ file format trimmed and filtered using FASTP as described above before being converted to BAM files.

Alignment of unmapped reads to microbial genomes

The unaligned reads were assigned to microbial taxa using PathSeq[18] v4.1.8.1 (http://software.broadinstitute.org/pathseq/) with the arguments "*--filter-duplicates false --min-score-identity .7*". We constructed the reference microbial genome database by downloading the set of complete viral, bacterial and fungal genomes from RefSeq release 201[49]. We subsampled at least one genome from each species including any genomes annotated as either "reference genome" or "representative genome" as well as the genomes of the three *Salmonella* strains used in the analyzed datasets. To mitigate vector contamination, we identified regions of suspected vector contamination (including "weak" matches) in the genomes using Vecscreen_plus_taxonomy (https://github.com/aaschaffer/vecscreen_plus_taxonomy) with the UniVec Database (ftp://ftp.ncbi.nlm.nih.gov/pub/UniVec/.) and filtered any reads that aligned to these regions[50]. The alternative microbial read identification on the *Salmonella* datasets was performed using CAMMiQ v0.1 (https://github.com/algo-cancer/CAMMiQ) using only the above set of genomes aggregated at the genus level.

Differential abundance quantification

We define the abundance of a particular microbe in each cell to be the number of unambiguous reads assigned to the relevant genome(s) by PathSeq The abundances are normalized using the computeSpikeFactors function from scran[51] v1.16.0 (https://github.com/MarioniLab/scran), which computes the library size factors using the sum of the spike-in sequences. To limit the number of hypotheses, we only test microbial taxa with counts per million microbial reads > 10 in at least 50% of the cells from a cell-type. The logged normalized read counts are compared across cell-types using the findMarkers function from scran v1.16.0 with arguments "*test='wilcox', lfc=0.5, block='plate'*". The findMarkers function from scran v1.16.0 makes inconsistent assumptions about how to distribute the values of the null distribution depending on whether the user specifies *"direction = 'up'"* or *"direction = 'down'"* (a one-sided test) or the user specifies *"direction = 'any'"* when the parameter *lfc* is greater than zero (https://github.com/MarioniLab/scran/issues/86). The assumption for the one-sided test models our intent so we ran the comparison twice, once using with "*direction='up'"* and once with "*direction='down'"*, selected the result with the smaller p-value for each microbial taxa and converted the one-sided p-value to the two-sided p-value by taking the minimum of 1 and 2*p-value as described in a standard reference[52] (page 79).

Differential presence quantification

We define the presence of a particular microbial taxon in each cell to be 1 if there are > 0 UMIs assigned unambiguously to the relevant genome(s) by PathSeq. We only analyze microbial taxa that are present in at least five cells total and at least 1% of the cells of a cell-type. We compare the presence of microbial taxa between two cell-types using the fisher.test implementation in R

of Fisher's exact test with default parameters, which reports the two-sided p-value and the conditional Maximum Likelihood Estimate (MLE).

False discovery rate correction

We used two different approaches for correcting p-values for multiple hypotheses. For the CSI-Microbes results from the *Salmonella* dataset, we run CSI-Microbes separately for each taxonomic level and correct for the number of taxa tested at that taxonomic level using the Benjamini-Hochberg procedure[53]. For the CSI-Microbes results from the cancer datasets, we leveraged the finding from the *Salmonella* dataset that CSI-Microbes can detect differentially abundant classes. For each class, we construct the taxonomic tree using RefSeq v201 and calculate the FDR for members of that class using the hFDR.adjust function from the structSSI package[54] (https://github.com/cran/structSSI). hFDR.adjust implements the "outer-nodes" method of Yekutieli[27], which is the method from that paper that is theoretically best suited for testing parent-child taxa in a taxonomic tree. To account for the multiple class hypotheses, we multiply the class-specific hFDR by the number of classes analyzed by CSI-Microbes to give the overall hierarchical FDR (hFDR). We compared the hFDR approach described above with FDR correction at the species level for the differential abundance of *Salmonella enterica* in the *Salmonella* Smart-seq2 dataset and find that the hFDR approach reports a more significant FDR-corrected p-value than the species-corrected FDR approach (1.58e-8 vs. 2.54e-8).

Normalization model

We extend the model used by decontam[22] to include host and spike-in sequences such that we let the total sample RNA (T) be a mixture of 3 components: human RNA (H), spike-in RNA (S) and

microbial RNA (M). We can further divide the microbial RNA into contaminating microbial RNA (cM) and true microbial RNA (tM). One previously observed pattern of contaminants is the frequency of contaminating microbial RNA (cM) is likely to be inversely correlated with the human RNA concentration[22]. We note that the frequency of spike-in RNA is also likely to be inversely correlated with the human RNA concentration and therefore the frequency of spike-in RNA should be correlated with the frequency of contaminant RNA. Therefore, spike-in based normalization should remove any differences in the frequency of contaminating sequences between cells.

Comparison to 16S tumor microbiome findings

We compared our findings of presence of bacterial taxa as numerical identifiers in NCBI's Taxonomy tree[28] to previously published findings[10]. To do this comparison, we had to i) map the published findings to numerical taxa and to assess which of the taxa they found are in our reference database. One of the key advantages of their 16S method is that it can find taxa for which there is no complete genome. In principle, CSI-Microbes can also use sub-genomic sequences in the reference database, but we chose not to use partial genomes.

In the previous study[10], microbial species were presented by name, which can lead to ambiguities because there are many synonyms and the preferred genus-species name may change over time. We were able to identify NCBI Taxonomy IDs for 1,783 of the identified species. 739 of these 1,783 species have at least one completely sequenced genome and were included in our microbial database. These species included 17 of the species reported to be enriched in lung cancer[10].

Differential expression between infected and uninfected tumor cells

We start with four groups of infected tumor cells (one group per patient) and seven groups of uninfected tumor cells (one group per patient except for TH225 and TH226, which we exclude because they have only ten and eleven cancer cells respectively). We first identified genes that are differentially expressed with an average minimum $\log_2$ fold-change=.5 between one group of infected tumor cells and one group of uninfected tumor cells, which provides 28 pair-wise comparisons (these comparisons are independent because we use a non-overlapping subset of uninfected tumor cells from each patient). We calculate a set of differentially expressed genes for each group of infected tumor cells by taking the median value of the Holm-corrected p-value for each gene (using the argument 'pval.type="some"' for the function findMarkers from scran[51]). Finally, the differentially expressed genes for tumor cells from each infected tumor are combined using Stouffer's Z score method[21]. We follow the procedure described earlier for computing the two-sided p-value.

Gene set enrichment analysis

We performed pathway enrichment analysis and visualization as previously suggested[55]. To perform GSEA between the infected and uninfected tumor cells, we first performed differential expression analysis as described above except that we used LFC=0 to limit the number of genes with p-value=1 and thereby the number of tied genes. Next, we ranked genes by multiplying the -$\log_{10}$(p-value) by -1 (AUC > 0.50 for Wilcoxon rank sum test) or 1 (AUC ≤ .50). Finally, we performed gene set enrichment analysis using the ranked genes list and the GSEAPreranked function of the GSEA tool[56] v4.1.0 (http://www.gsea-msigdb.org/gsea/index.jsp) with default settings and seed=149 with the gene ontology biological processes gene set from the molecular

signature database (MSigDB)[56–59] v7.3 (http://www.gsea-msigdb.org/gsea/msigdb/index.jsp). We visualized the enriched gene sets using Enrichment Map[38] v3.3.1 (https://enrichmentmap.readthedocs.io/) with parameters: FDR q-value < .02 (node cutoff) and Jacard Overlap Combined Index (k constant=0.5) > .375 (edge cutoff). Clustering was performed on the graph using MCL Cluster from clusterMaker2[60] via AutoAnnotate[61] and annotated using WordCloud: Adjacent Words from AutoAnnotate[61]. Annotations were manually reviewed and edited where appropriate.

Acknowledgements

This research was supported in part by the Intramural Research Program of the National Institutes of Health, National Cancer Institute. WR's contribution to this research was supported in part by NSF award DGE-1632976. This work utilized the computational resources of the NIH HPC Biowulf cluster. (http://hpc.nih.gov). The authors thank Wolfgang Resch for assistance running this analysis on the NIH HPC Biowulf cluster. The authors thank Anna Aulicino, Noa Bossel Ben-Moshe, Leigh Greathouse, Hector Corrado-Bravo, Norma Andrews, Steve Christensen, Itay Tirosh, Livnat Jerby-Arnon, Yunhua Zhu, Chen Zhao, Spyros Darmanis, Frank Lowery and Sri Krishna for useful discussions about this project.

Author Contributions

Conceptualization of the project - WR AAS ER. Data curation - WR JSL KZ AAS. Formal analysis - WR. Methods - WR FS MDML RP AAS ER. Software development - WR FS EMG KZ. Software testing and documentation - WR FS AAS. Visualization - WR FS. Writing original draft - WR ER. Writing, reviewing, and editing of later drafts - WR FS EMG MDML AAS ER. Supervision - SCS AAS ER.

Competing Interests

RP is a co-founder of Ocean Genomics, Inc. The other authors declare that they have no competing interests.

Supplementary Tables

Supplementary Table 1: Results from both 10x and Smart-seq2 *Salmonella* datasets including read counts of protein coding genes (minimum 5 reads) from *LT2* and *D23580* genomes and the output of CSI-Microbes for all microbial taxa from the 10x dataset (minimum 5 cells with at least one UMI) and the Smart-seq2 dataset (minimum 10 CPM in least 50% of either infected or bystander cells).

Supplementary Table 2: Output of CSI-Microbes between tumor and non-tumor cells on patients from the Merkel cell carcinoma 10x cohort[30].

Supplementary Table 3: Output of CSI-Microbes between tumor and non-tumor cells on patients from the colorectal carcinoma 10x cohort[31].

Supplementary Table 4: Output of CSI-Microbes between tumor and non-tumor cells on patients from the lung cancer Smart-seq2 cohort[33].

Supplementary Table 5: Complete list of differentially expressed genes between infected and uninfected lung tumor cells.

Supplementary Table 6: Up and down-regulated gene sets in infected tumor cells compared to uninfected tumor cells using GSEA.

References

1. Bullman, S. *et al.* Analysis of Fusobacterium persistence and antibiotic response in colorectal cancer. *Science (80-.).* **358**, 1443–1448 (2017).
2. Castellarin, M. *et al.* Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Res.* **22**, 299–306 (2012).
3. Gur, C. *et al.* Binding of the Fap2 protein of fusobacterium nucleatum to human inhibitory receptor TIGIT protects tumors from immune cell attack. *Immunity* **42**, 344–355 (2015).

4. Gur, C. *et al.* Fusobacterium nucleatum supresses anti-tumor immunity by activating CEACAM1. *Oncoimmunology* **8**, e1581531 (2019).
5. Kostic, A. D. *et al.* Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Res.* **22**, 292–298 (2012).
6. Yu, T. C. *et al.* Fusobacterium nucleatum Promotes Chemoresistance to Colorectal Cancer by Modulating Autophagy. *Cell* **170**, 548-563.e16 (2017).
7. Pleguezuelos-Manzano, C. *et al.* Mutational signature in colorectal cancer caused by genotoxic pks + E. coli. *Nature* **580**, 269–273 (2020).
8. Geller, L. T. *et al.* Potential role of intratumor bacteria in mediating tumor resistance to the chemotherapeutic drug gemcitabine. *Science (80-.).* **1160**, 1156–1160 (2017).
9. Poore, G. D. *et al.* Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* **579**, 567–574 (2020).
10. Nejman, D. *et al.* The human tumor microbiome is composed of tumor type-specific intracellular bacteria. *Science (80-.).* **368**, 973–980 (2020).
11. Abed, J. *et al.* Fap2 Mediates Fusobacterium nucleatum Colorectal Adenocarcinoma Enrichment by Binding to Tumor-Expressed Gal-GalNAc. *Cell Host Microbe* **20**, 215–225 (2016).
12. Kalaora, S. *et al.* Identification of bacteria-derived HLA-bound peptides in melanoma. *Nature* (2021) doi:10.1038/s41586-021-03368-8.
13. Weller, R. & Ward, D. M. Selective recovery of 16S rRNA sequences from natural microbial communities in the form of cDNA. *Appl. Environ. Microbiol.* **55**, 1818–1822 (1989).
14. Fox, G. E., Pechman, K. R. & Woese, C. R. Comparative cataloging of 16S ribosomal ribonucleic acid: molecular approach to procaryotic systematics. *Int. J. Syst. Bacteriol.* **27**, 44–57 (1977).
15. Steuerman, Y. *et al.* Dissection of Influenza Infection In Vivo by Single-Cell RNA Sequencing. *Cell Syst.* **6**, 679-691.e4 (2018).
16. Bost, P. *et al.* Host-Viral Infection Maps Reveal Signatures of Severe COVID-19 Patients. *Cell* **181**, 1475-1488.e12 (2020).
17. Avital, G. *et al.* scDual-Seq: mapping the gene regulatory program of Salmonella infection by host and pathogen single-cell RNA-sequencing. *Genome Biol.* **18**, 200

(2017).

18. Walker, M. A. *et al.* GATK PathSeq: a customizable computational tool for the discovery and identification of microbial sequences in libraries from eukaryotic hosts. *Bioinformatics* **34**, 4287–4289 (2018).
19. Hanley, J. A. & McNeil, B. J. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology* **143**, 29–36 (1982).
20. Fisher, R. A. *Statistical methods for research workers*. *Statistical methods for research workers, 5th ed.* (Oliver and Boyd, 1934).
21. Stouffer, S. A., Suchman, E. A., DeVinney, L. C. & Williams, R. M. *The American Soldier. Adjustment During Army LIfe*. (Princeton University Press, 1949).
22. Davis, N. M., Proctor, Di. M., Holmes, S. P., Relman, D. A. & Callahan, B. J. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* **6**, 226 (2018).
23. Aulicino, A. *et al.* Invasive Salmonella exploits divergent immune evasion strategies in infected and bystander dendritic cell subsets. *Nat. Commun.* **9**, 4883 (2018).
24. Bossel Ben-Moshe, N. *et al.* Predicting bacterial infection outcomes using single cell RNA-sequencing analysis of human immune cells. *Nat. Commun.* **10**, 3266 (2019).
25. Zhu, K. *et al.* Strain Level Microbial Detection and Quantification with Applications to Single Cell Metagenomics. *bioRxiv* (2020).
26. Morgulis, A. & Agarwala, R. SRPRISM (Single Read Paired Read Indel Substitution Minimizer): an efficient aligner for assemblies with explicit guarantees. *Gigascience* **9**, (2020).
27. Yekutieli, D. Hierarchical false discovery rate-controlling methodology. *J. Am. Stat. Assoc.* **103**, 309–316 (2008).
28. Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Res.* **40**, D136-143 (2012).
29. Feng, H., Shuda, M., Chang, Y. & Moore, P. S. Clonal integration of a polyomavirus in human Merkel cell carcinoma. *Science (80-.).* **319**, 1096–1100 (2008).
30. Paulson, K. G. *et al.* Acquired cancer resistance to combination immunotherapy from transcriptional loss of class I HLA. *Nat. Commun.* **9**, (2018).
31. Lee, H. O. *et al.* Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nat. Genet.* **52**, 594–603 (2020).

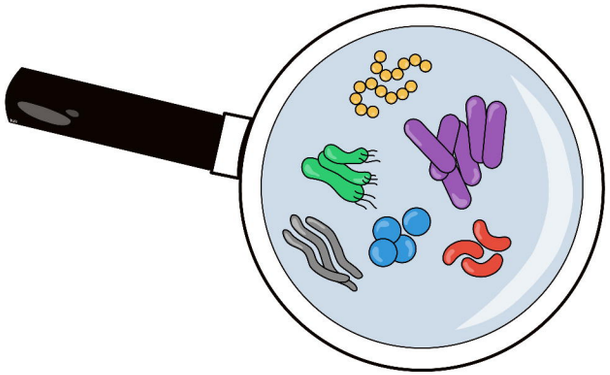
32. Kleessen, B., Kroesen, A. J., Buhr, H. J. & Blaut, M. Mucosal and invading bacteria in patients with inflammatory bowel disease compared with controls. *Scand. J. Gastroenterol.* **37**, 1034–1041 (2009).
33. Maynard, A. *et al.* Therapy-Induced Evolution of Human Lung Cancer Revealed by Single-Cell RNA Sequencing. *Cell* **182**, 1232-1251.e22 (2020).
34. Mayslich, C., Grange, P. A. & Dupin, N. Cutibacterium acnes as an opportunistic pathogen: An update of its virulence-associated factors. *Microorganisms* **9**, 1–21 (2021).
35. Bae, Y. *et al.* Intracellular propionibacterium acnes infection in glandular epithelium and stromal macrophages of the prostate with or without cancer. *PLoS One* **9**, e90324 (2014).
36. Eishi, Y. Etiologic link between sarcoidosis and Propionibacterium acnes. *Respir. Investig.* **51**, 56–68 (2013).
37. Brophy, M. B. & Nolan, E. M. Manganese and microbial pathogenesis: Sequestration by the mammalian immune system and utilization by microorganisms. *ACS Chem. Biol.* **10**, 641–651 (2015).
38. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation. *PLoS One* **5**, e13984 (2010).
39. Radhakrishnan, G. K. & Splitter, G. A. Modulation of host microtubule dynamics by pathogenic bacteria. *Biomol. Concepts* **3**, 571–580 (2012).
40. Wang, F. *et al.* RNAscope: A novel in situ RNA analysis platform for formalin-fixed, paraffin-embedded tissues. *J. Mol. Diagnostics* **14**, 22–29 (2012).
41. Silva, M. T. Classical labeling of bacterial pathogens according to their lifestyle in the host: Inconsistencies and alternatives. *Front. Microbiol.* **3**, 71 (2012).
42. Mohanty, B. K. & Kushner, S. R. Bacterial/archaeal/organellar polyadenylation. *Wiley Interdiscip Rev RNA* **2**, 256–276 (2010).
43. Cao, G. J. & Sarkar, N. Identification of the gene for an Escherichia coli poly(A) polymerase. *Proc. Natl. Acad. Sci. U. S. A.* **89**, (1992).
44. Régnier, P. & Marujo, P. E. Polyadenylation and Degradation of RNA in Prokaryotes. in *Madame Curie Bioscience Database [Internet]* (Landes Bioscience, 2013).
45. Köster, J. & Rahmann, S. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).

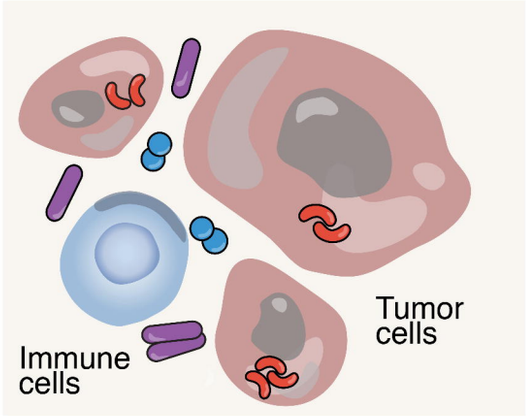
46. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
47. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
48. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
49. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
50. Schäffer, A. A. *et al.* VecScreen-plus-taxonomy: Imposing a tax(onomy) increase on vector contamination screening. *Bioinformatics* **34**, 755–759 (2018).
51. Lun, A. T. L., Mccarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor [version 2 ; referees : 3 approved , 2 approved with reservations]. *F1000Research* **5**, (2016).
52. Sokal, R. R. & Rohlf, F. J. *Biometry*. (W. H. Freeman, 1995).
53. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
54. Sankaran, K. & Holmes, S. structSSI: Simultaneous and selective inference for grouped or hierarchically structured data. *J. Stat. Softw.* **59**, 1–21 (2014).
55. Reimand, J. *et al.* Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.* **22**, 924–934 (2019).
56. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S. & Ebert, B. L. Gene set enrichment analysis : A knowledge-based approach for interpreting genome-wide. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).
57. Ashburner, M. *et al.* Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
58. Carbon, S. *et al.* The Gene Ontology resource: Enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334 (2021).
59. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
60. Morris, J. H. *et al.* ClusterMaker: A multi-algorithm clustering plugin for Cytoscape.

BMC Bioinformatics **12**, (2011).

61. Kucera, M., Isserlin, R., Arkhangorodsky, A. & Bader, G. D. AutoAnnotate: A Cytoscape app for summarizing networks with semantic annotations [version 1; peer review: 2 approved]. *F1000Research* **5**, (2016).

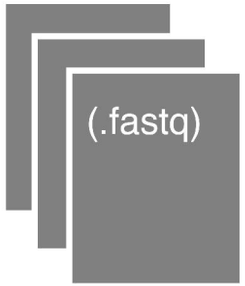
CSI-Microbes





Input

Sequencing read files



Metadata

Celltype	Plate	Cell
Tumor	P1	C1
Tumor	P1	C2
Immune	P1	C3
Immune	P1	C4
...	...	...

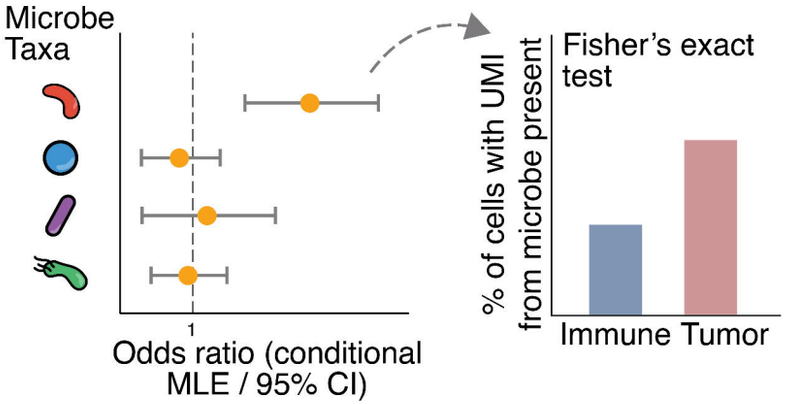
Identification

Map reads against

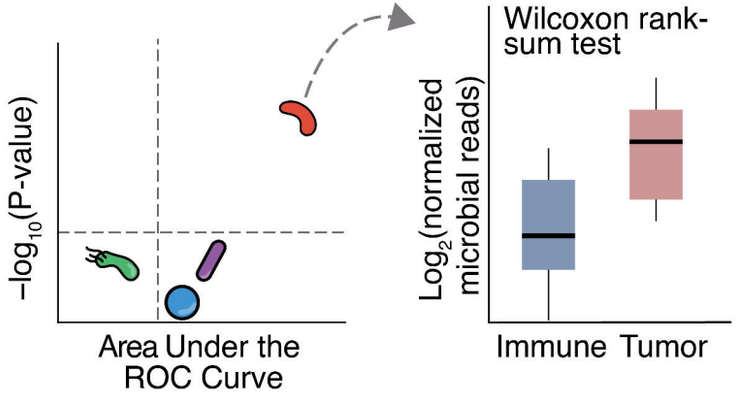
Spike-ins
(Smart-seq2 only)

Analysis

Differential Presence (10x)

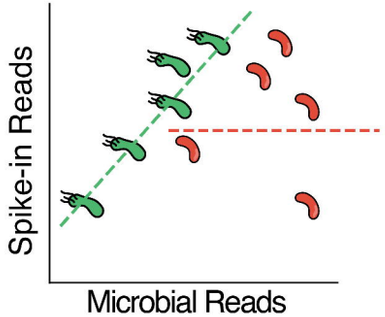


Differential Abundance (Smart-seq2)

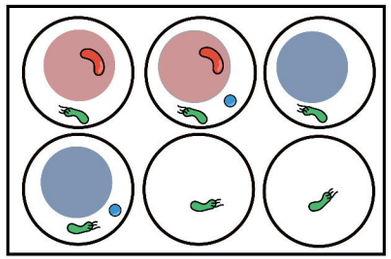


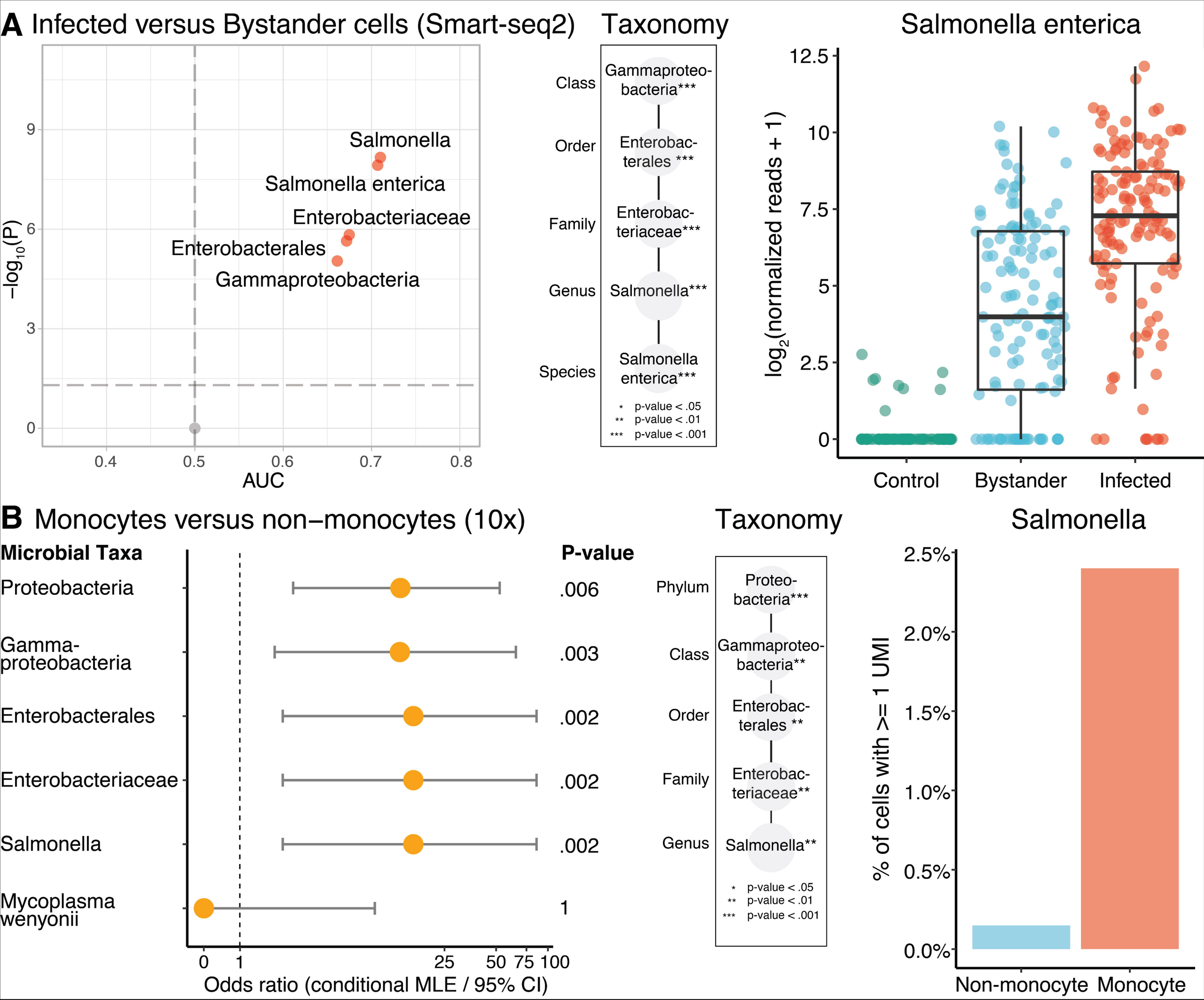
Validation

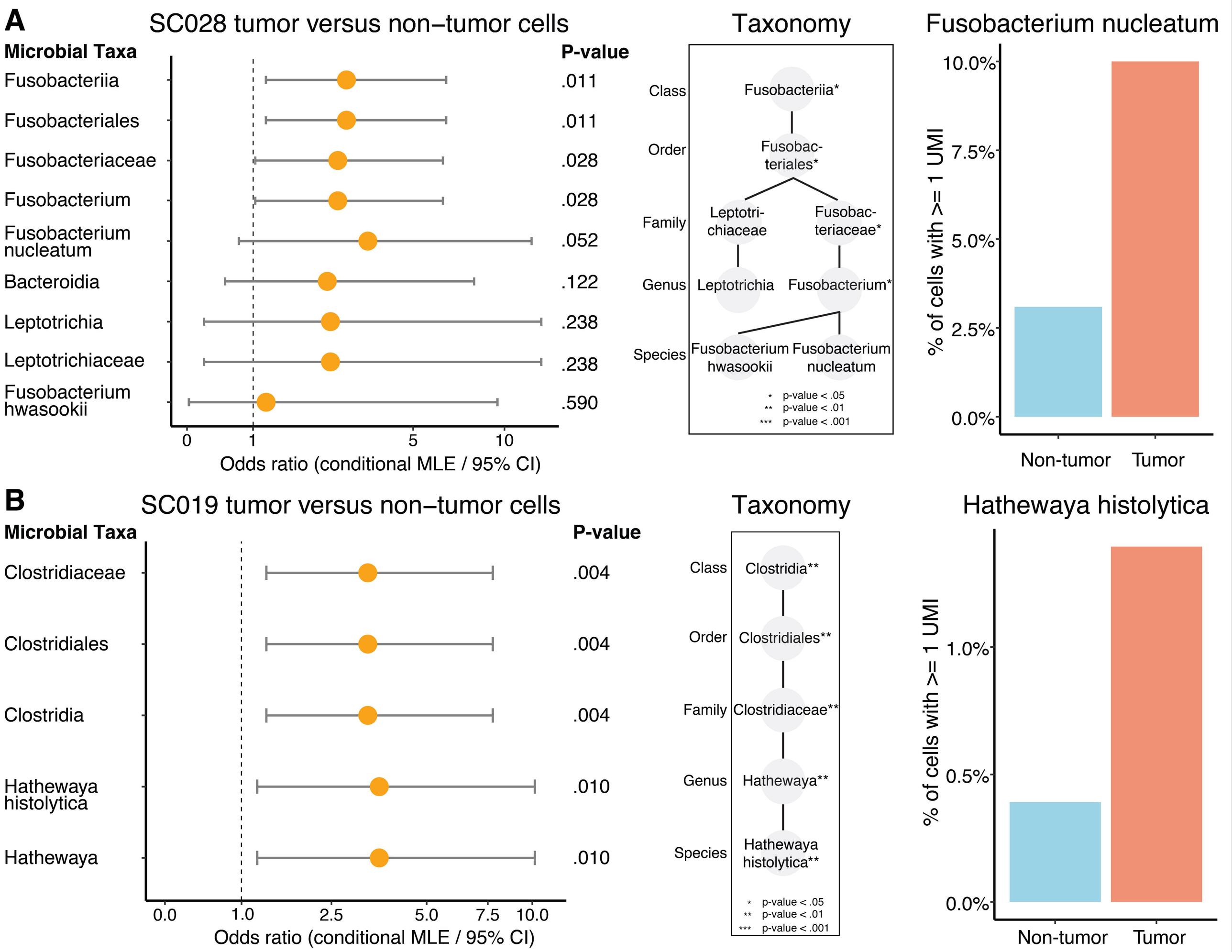
Spike-in Test

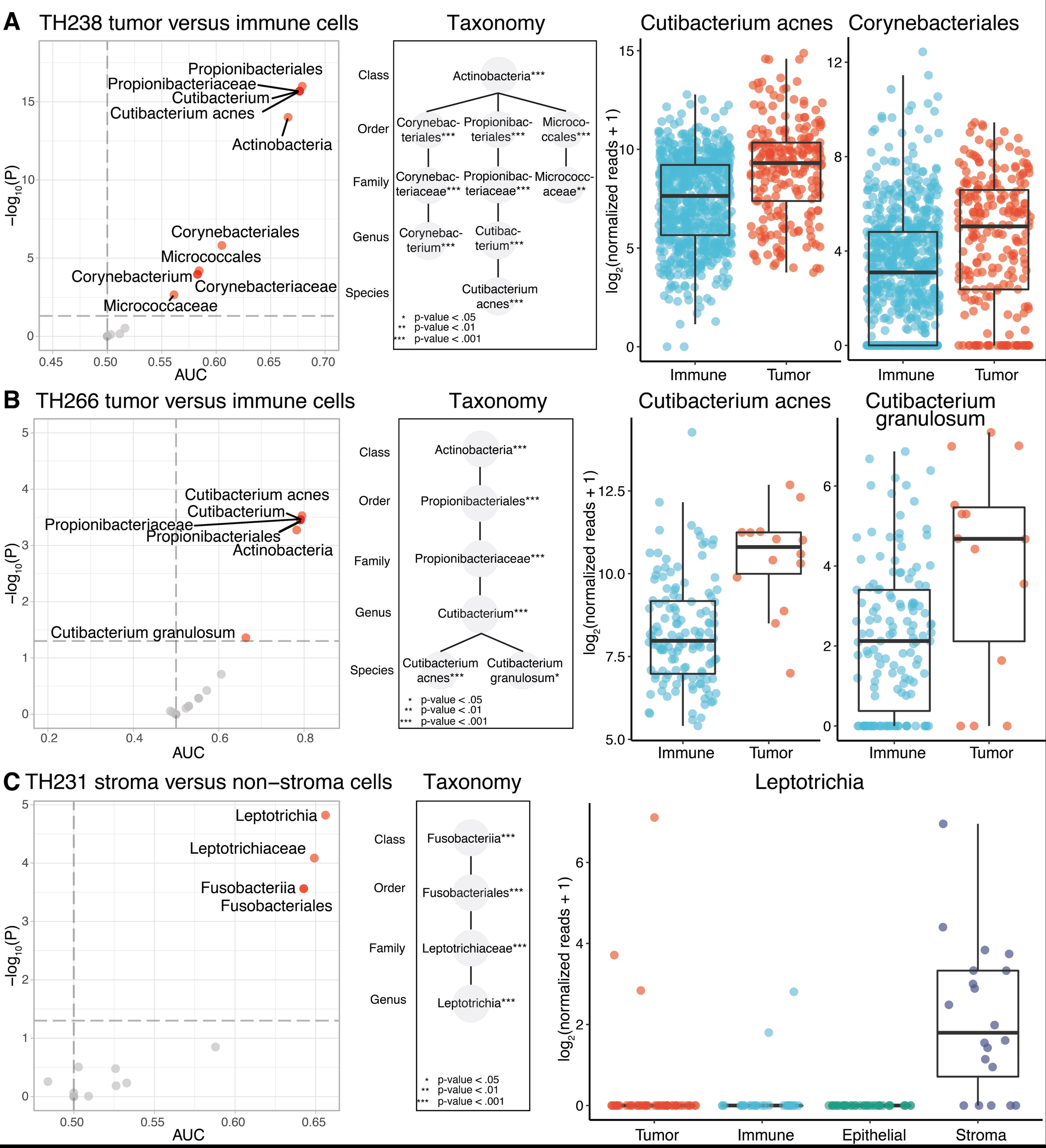


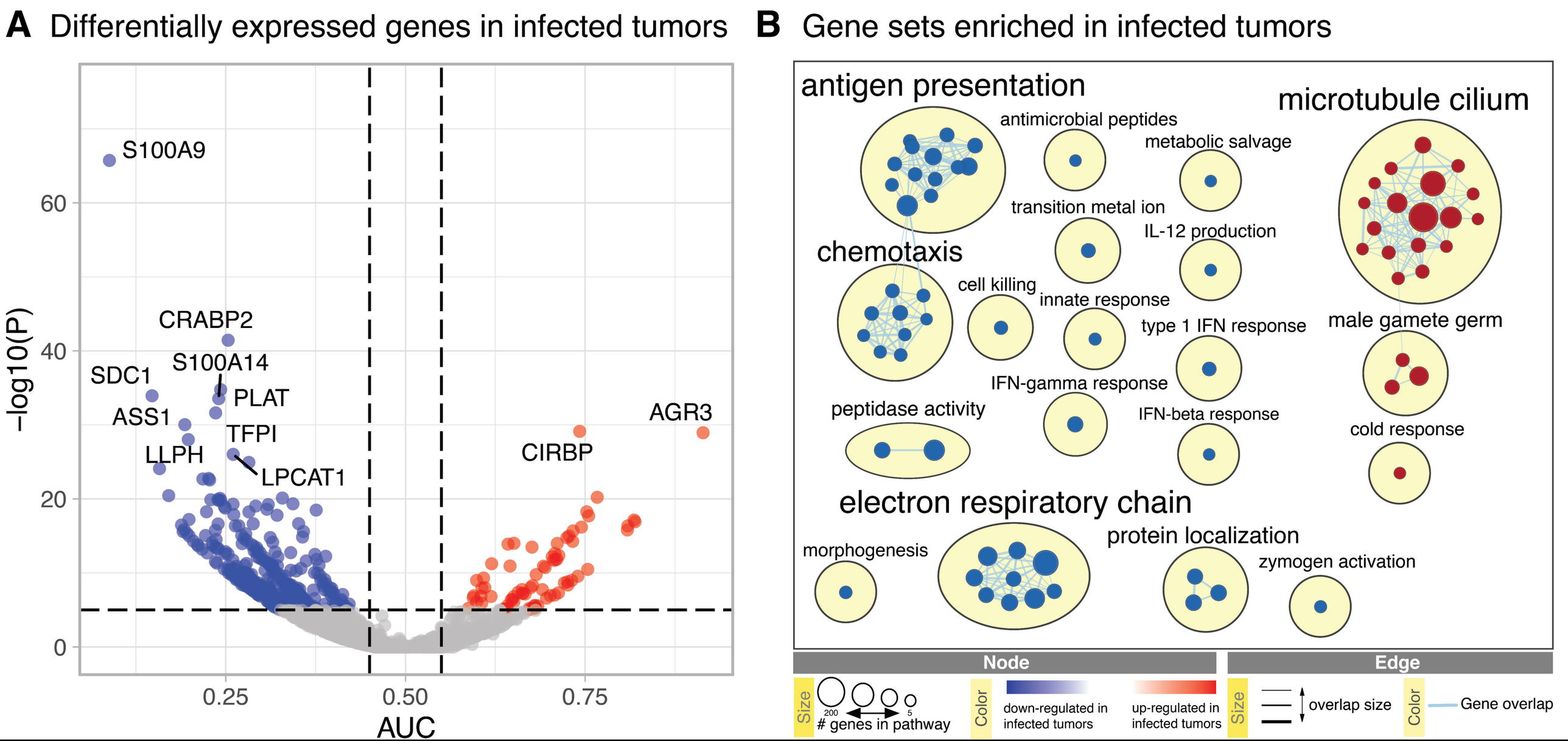
Empty Wells Test

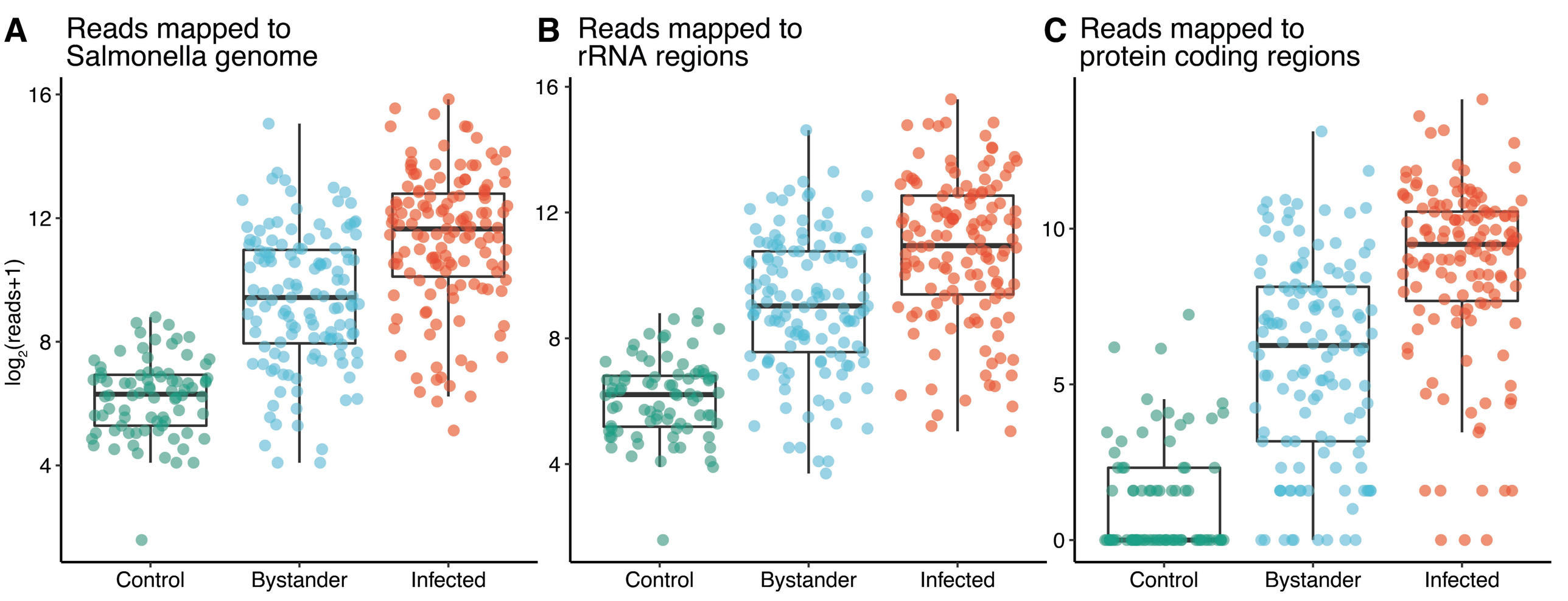


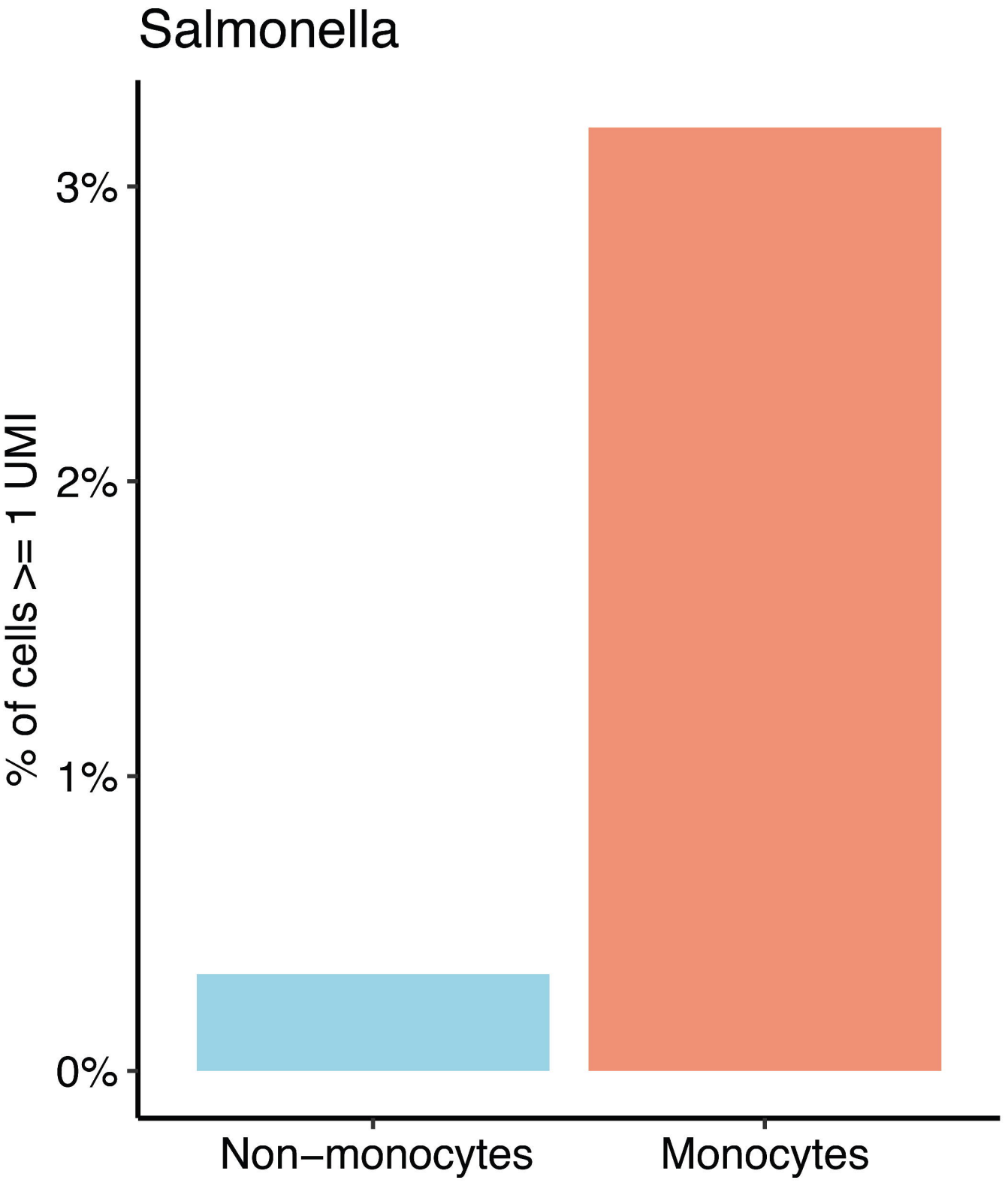

A Infected versus Bystander cells (Smart-seq2)
Salmonella
Salmonella enterica
Enterobacteriaceae
Enterobacterales
Gammaproteobacteria
−log10(P)
AUC
Taxonomy
Class
Gammaproteo-bacteria***
Order
Enterobac-terales ***
Family
Enterobac-teriaceae***
Genus
Salmonella***
Species
Salmonella enterica***
* p-value < .05
** p-value < .01
*** p-value < .001
Salmonella enterica
log2(normalized reads + 1)
Control
Bystander
Infected
B Monocytes versus non−monocytes (10x)
Microbial Taxa
P-value
Proteobacteria .006
Gamma-proteobacteria .003
Enterobacterales .002
Enterobacteriaceae .002
Salmonella .002
Mycoplasma wenyonii 1
Odds ratio (conditional MLE / 95% CI)
Taxonomy
Phylum
Proteo-bacteria***
Class
Gammaproteo-bacteria**
Order
Enterobac-terales **
Family
Enterobac-teriaceae**
Genus
Salmonella**
* p-value < .05
** p-value < .01
*** p-value < .001
Salmonella
% of cells with >= 1 UMI
Non-monocyte
Monocyte

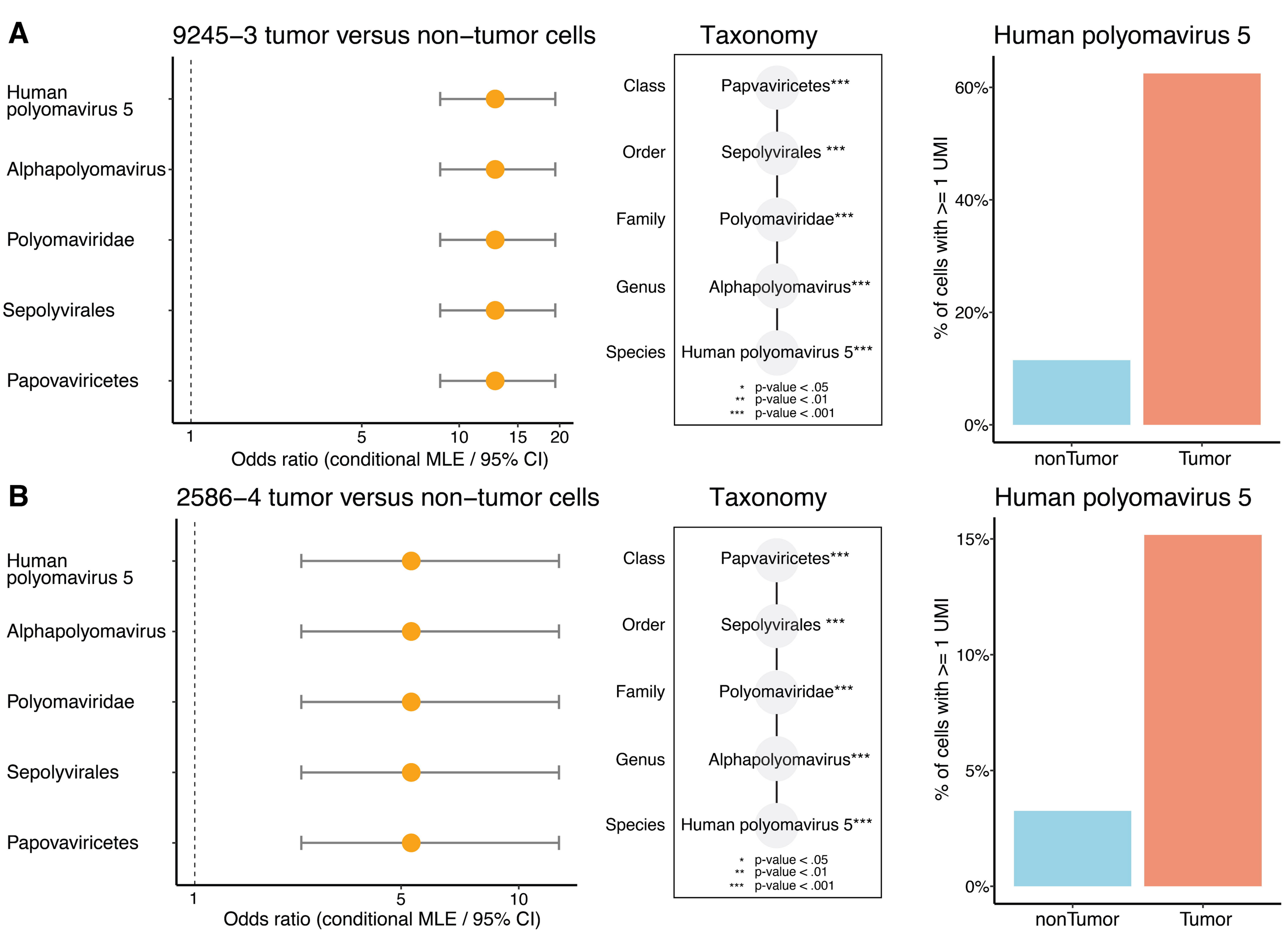

A
SC028 tumor versus non−tumor cells
Microbial Taxa
P-value
Fusobacteriia .011
Fusobacteriales .011
Fusobacteriaceae .028
Fusobacterium .028
Fusobacterium nucleatum .052
Bacteroidia .122
Leptotrichia .238
Leptotrichiaceae .238
Fusobacterium hwasookii .590
0 1 5 10
Odds ratio (conditional MLE / 95% CI)
Taxonomy
Class Fusobacteriia*
Order Fusobac-teriales*
Family Leptotri-chiaceae Fusobac-teriaceae*
Genus Leptotrichia Fusobacterium*
Species Fusobacterium hwasookii Fusobacterium nucleatum
* p-value < .05
** p-value < .01
*** p-value < .001
Fusobacterium nucleatum
% of cells with >= 1 UMI
10.0% 7.5% 5.0% 2.5% 0.0%
Non-tumor Tumor
B
SC019 tumor versus non−tumor cells
Microbial Taxa
P-value
Clostridiaceae .004
Clostridiales .004
Clostridia .004
Hathewaya histolytica .010
Hathewaya .010
0.0 1.0 2.5 5.0 7.5 10.0
Odds ratio (conditional MLE / 95% CI)
Taxonomy
Class Clostridia**
Order Clostridiales**
Family Clostridiaceae**
Genus Hathewaya**
Species Hathewaya histolytica**
* p-value < .05
** p-value < .01
*** p-value < .001
Hathewaya histolytica
% of cells with >= 1 UMI
1.0% 0.5% 0.0%
Non-tumor Tumor


A
TH238 tumor versus immune cells
Propionibacteriales
Propionibacteriaceae
Cutibacterium
Cutibacterium acnes
Actinobacteria
Corynebacteriales
Micrococcales
Corynebacterium
Corynebacteriaceae
Micrococcaceae
−log10(P)
AUC
Taxonomy
Class
Order
Family
Genus
Species
Actinobacteria***
Corynebac-teriales***
Propionibac-teriales***
Micrococ-cales***
Corynebac-teriaceae***
Propionibac-teriaceae***
Micrococc-aceae**
Corynebac-terium***
Cutibac-terium***
Cutibacterium acnes***
* p-value < .05
** p-value < .01
*** p-value < .001
Cutibacterium acnes
Corynebacteriales
log2(normalized reads + 1)
Immune
Tumor
B
TH266 tumor versus immune cells
Cutibacterium acnes
Cutibacterium
Propionibacteriaceae
Propionibacteriales
Actinobacteria
Cutibacterium granulosum
Actinobacteria***
Propionibacteriales***
Propionibacteriaceae***
Cutibacterium***
Cutibacterium acnes***
Cutibacterium granulosum*
Cutibacterium granulosum
C
TH231 stroma versus non−stroma cells
Leptotrichia
Leptotrichiaceae
Fusobacteriia
Fusobacteriales
Fusobacteriia***
Fusobacteriales***
Leptotrichiaceae***
Leptotrichia***
Tumor
Immune
Epithelial
Stroma


A Differentially expressed genes in infected tumors
S100A9
CRABP2
S100A14
SDC1
PLAT
ASS1
TFPI
LLPH
LPCAT1
CIRBP
AGR3
−log10(P)
60
40
20
0
0.25
0.50
0.75
AUC
B Gene sets enriched in infected tumors
antigen presentation
antimicrobial peptides
metabolic salvage
microtubule cilium
transition metal ion
IL-12 production
chemotaxis
cell killing
innate response
type 1 IFN response
male gamete germ
IFN-gamma response
peptidase activity
IFN-beta response
cold response
electron respiratory chain
protein localization
morphogenesis
zymogen activation
Node
Edge
Size
200
5
genes in pathway
Color
down-regulated in infected tumors
up-regulated in infected tumors
Size
overlap size
Color
Gene overlap


A
Reads mapped to Salmonella genome
B
Reads mapped to rRNA regions
C
Reads mapped to protein coding regions
log2(reads+1)
16
12
8
4
10
5
0
Control
Bystander
Infected


Salmonella
% of cells >= 1 UMI
3%
2%
1%
0%
Non−monocytes
Monocytes


A
9245–3 tumor versus non–tumor cells
Human polyomavirus 5
Alphapolyomavirus
Polyomaviridae
Sepolyvirales
Papovaviricetes
1
5
10
15
20
Odds ratio (conditional MLE / 95% CI)
Taxonomy
Class
Papvaviricetes***
Order
Sepolyvirales ***
Family
Polyomaviridae***
Genus
Alphapolyomavirus***
Species
Human polyomavirus 5***
* p-value < .05
** p-value < .01
*** p-value < .001
Human polyomavirus 5
% of cells with >= 1 UMI
60%
40%
20%
0%
nonTumor
Tumor
B
2586–4 tumor versus non–tumor cells
Human polyomavirus 5
Alphapolyomavirus
Polyomaviridae
Sepolyvirales
Papovaviricetes
1
5
10
Odds ratio (conditional MLE / 95% CI)
Taxonomy
Class
Papvaviricetes***
Order
Sepolyvirales ***
Family
Polyomaviridae***
Genus
Alphapolyomavirus***
Species
Human polyomavirus 5***
* p-value < .05
** p-value < .01
*** p-value < .001
Human polyomavirus 5
% of cells with >= 1 UMI
15%
10%
5%
0%
nonTumor
Tumor