Identifying the Landscape of Intratumoral Microbes via a Single Cell Transcriptomic Analysis

Welles Robinson[1,2], Fiorella Schischlik[1], E. Michael Gertz[1], 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

Abstract

Microbial taxa that are differentially abundant between cell types are likely to be intracellular. Here we describe a new computational pipeline called CSI-Microbes (computational identification of **C**ell type **S**pecific **I**ntracellular **Microbes**) that aims to identify such putative intracellular species from single cell RNA-seq data in a given tumor sample. CSI-microbes also includes additional steps that can be applied to filter out microbial contaminants from the *bona fide* microbial residents of cells in the patients. We first test and validate CSI-microbes on a dataset of immune cells deliberately infected with *Salmonella*. We then apply CSI-microbes to identify intracellular microbes in breast cancer and melanoma. We identify *Streptomyces* as differentially abundant in the tumor cells of one breast cancer sample. We further identify three bacterial genera and four fungal genera that are differentially abundant and hence likely to be intracellular in the tumor cells in melanoma samples. No cell type specific bacteria were identified in our analysis of brain tumor samples. In sum, CSI-Microbes offers a new way to identify likely intracellular microbes living within specific cell populations in malignant tumors, markedly extending upon previous studies aimed at inferring microbial abundance from bulk tumor expression data.

Introduction

Several recent papers have examined the tumor microbiome, pointing to its functional importance. 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 (Bullman et al., 2017; Castellarin et al. 2012; Gur et al., 2015, 2019; Kostic et al., 2012; Yu et al., 2017). *pks+ E. coli* have been shown to induce a mutation signature frequently found in colorectal carcinoma (Pleguezuelos-Manzano et al., 2020). In pancreatic cancer, a subset of Gammaproteobacteria were shown to mediate tumor resistance to chemotherapy (Geller et al., 2017). Recently, an analysis of the Cancer Genome Atlas (TCGA) cohort identified a variety of bacterial genera that reside in different tumor types, demonstrating that after filtering out potentially contaminant species, one can successfully build a predictor of cancer type based on tumors' microbial composition (Poore et al., 2020).

Despite these recent advances, many unresolved questions concerning the tumor microbiome remain. One major concern is the potential that microbial sequences identified

through next-generation sequencing may result from contamination accompanying the acquisition of the tumor samples and their sequencing, instead of being truly present in the tumor. A second unresolved question is whether the microbes identified in these bulk RNA and DNA sequencing studies exist extracellularly or intracellularly. Thirdly, even if the microbes identified are intracellular, the question remains whether they specifically reside in tumor cells, or in different types of immune or stromal cells.

Here, we present a computational approach named **CSI-Microbes** (computational identification of **C**ell type **S**pecific **I**ntracellular **Microbes**) that addresses these three research questions by identifying intracellular microbes that are *cell type-specific* while filtering out contaminating microbes. It is based on identifying microbial sequences from single-cell RNA-seq (scRNA-seq) datasets, such that the taxa containing those sequences are differentially abundant between distinct cell types. To the best of our knowledge, this is the first approach to identify microbial sequences from large-scale scRNA-seq cancer experiments. CSI-Microbes capitalizes on several aspects of scRNA-seq, including the use of spike-in RNAs, analyzing the contents of empty wells and examining the content of multiple cells from a given tumor. We first test and validate our approach on non-cancer data by demonstrating that CSI-Microbes identifies *Salmonella* as the only differentially abundant genus between monocyte-derived dendritic cells (moDCs) infected with *Salmonella* and moDCs infected with a mock control in a scRNA-seq experiment. Next, we apply our approach to five scRNA-seq datasets spanning 47 cancer patients across three cancer types. We find bioinformatic evidence for multiple instances of cell type-specific intracellular microbes in three patients spanning two cancer types.

Results

The inputs to CSI-Microbes are FASTQ files from scRNA-seq experiments and cell metadata, including cell type annotations and known sources of contamination such as the sequencing plate. The output is a ranked list of likely cell type-specific intracellular microbes and their differential cell type specific abundance. The algorithm proceeds in the following steps (**Figure 1A** and see Methods for a detailed description): **(1)** The first step of CSI-Microbes extracts scRNA-seq reads that map to microbial genomes after filtering the host reads. **(2)** Next, CSI-Microbes normalizes and log-transforms the microbial reads so they can be compared across cells. **(3)** Thirdly, the normalized microbial read counts are compared across the cell types while controlling for known sources of contamination. **(4)** Finally, when present in the dataset analyzed, CSI-Microbes employs two further testing and validation approaches that mirror gold-standard bulk microbiome approaches for filtering contaminants: those include, **(4a)** determining whether microbes appear more abundant in low-concentration samples using spike-in RNAs and **(4b)** filtering out microbes that appear in higher frequency in empty wells, which effectively function as negative controls (Davis et al., 2018).

To test the feasibility of identifying microbial reads from scRNA-seq datasets, we applied CSI-Microbes to two "gold-standard" datasets where immune cells were infected *ex vivo* with the intracellular bacteria *Salmonella enterica* and subsequently sequenced using scRNA-seq (Aulicino et al., 2018; Ben-Moshe et al., 2019). In the first dataset, where ~7,000 peripheral blood mononuclear cells (PBMCs) were infected with *Salmonella enterica* serovar Typhimurium and sequenced using 10x 3' sequencing, we identified very few reads that mapped to the

Salmonella genus. In the second dataset, however, which consists of 373 monocyte-derived dendritic cells (moDCs) infected with either the D23580 or the LT2 strain of *Salmonella enterica* and sequenced using Smart-seq2 platform, we could successfully identify a large number of reads that mapped to the *Salmonella* genus. Although there are multiple differences between the experiments, we hypothesized that much of this difference can be attributed to the sequencing protocol because the plate-based Smart-seq2 has been shown to capture many more transcripts per cell than the droplet-based 3' 10x protocol (See et al., 2018). For this reason, we chose to focus on scRNA-seq datasets generated by protocols based on SMART sequencing such as Smart-seq2, Fluidigm C1 and SMARTer moving forward in our analysis.

Analyzing the moDCs dataset, we used annotations provided by the authors and categorized the cells into two groups: cells that were infected with *Salmonella* and cells that were mock-infected and sequenced as controls. Reassuringly, CSI-Microbes identifies that the abundance of *Salmonella* was substantially higher in the infected cells compared to the mock-infected controls (**Figure 1B**). We found multiple *Salmonella* reads in every confirmed infected cell. In addition to the reads assigned to the *Salmonella* genus, we identified a large number of reads uniquely assigned to other microbes at the genera level, which are likely contaminates. However, importantly, CSI-Microbes identified *Salmonella* as the only differentially abundant microbe between the infected and mock-infected cells, demonstrating its ability to identify the true intracellular organism and filter out likely contaminating species (**Figure 1C**). As expected, there is no correlation between the abundance of the true intracellular bacteria *Salmonella* and the RNA spike-in transcripts (Spearman rank-correlation rho=.08, FDR=.54).

Figure 1

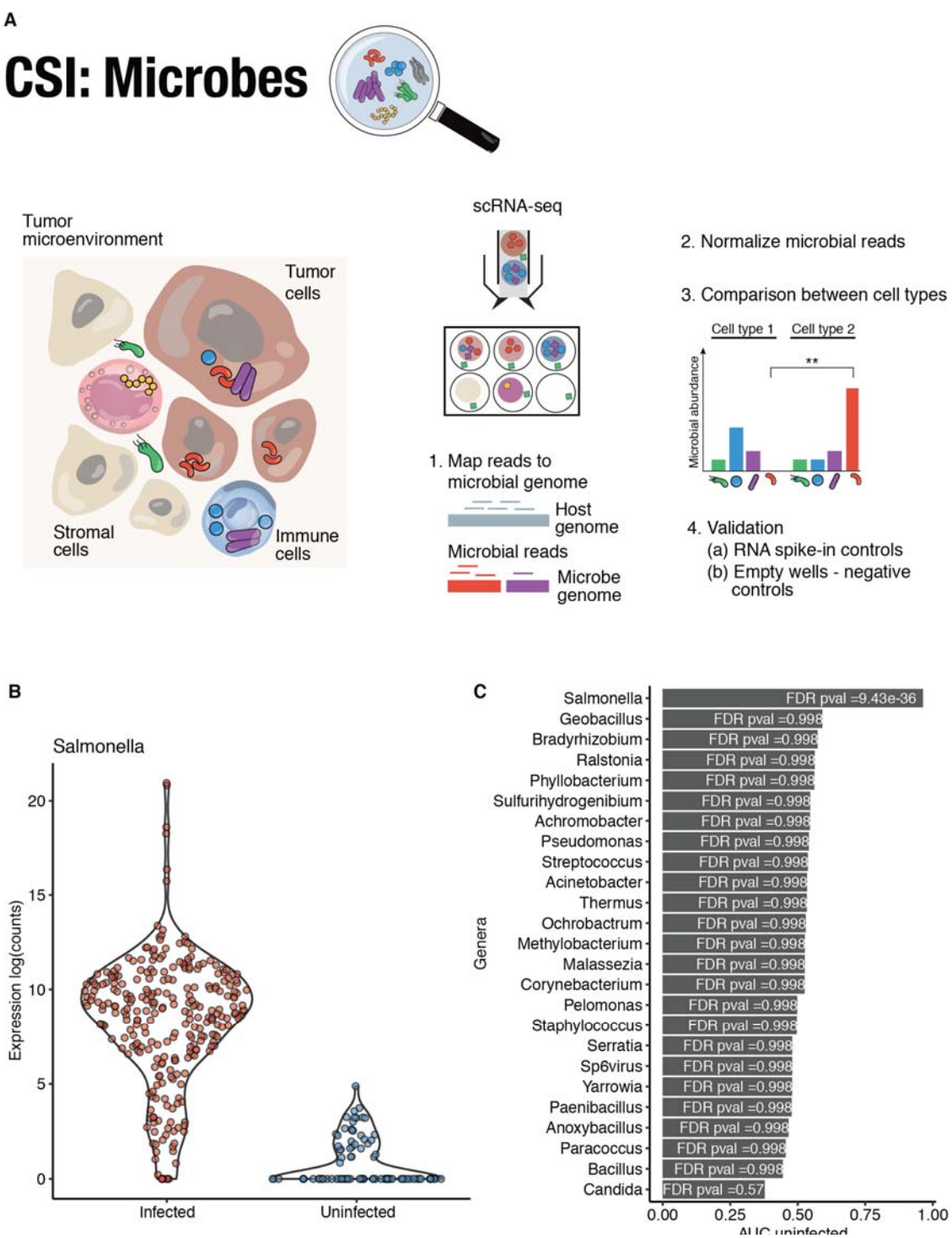

Figure 1. (A) Overview of the CSI-Microbes analysis pipeline. **(B)** Comparison of the normalized number of reads belonging to the *Salmonella* genus between infected and the mock-infected cells. **(C)** The 25 most differentially abundant genera (FDR) in this dataset according to CSI-Microbes ranked by their discriminatory power (AUC) between the infected and mock-infected cell types.

To study the landscape of intracellular bacteria in different tumor types, we analyzed five scRNA-seq datasets across three cancer types – breast cancer, brain cancer and melanoma – comprising ~14,000 single cells from 47 tumors sequenced using SMART-based protocols with full-length transcript capture. The two breast cancer datasets include 515 cells from 11 patients and ~1500 cells from six patients (Chung et al., 2017; Karaayvaz et al., 2018). The two brain cancer datasets consist of ~3500 cells from four glioma patients and ~6100 cells from 10 astrocytoma patients (Darmanis et al., 2017; Venteicher et al., 2017). The melanoma dataset contains ~2500 cells from 16 patients (Jerby-Arnon et al., 2018).

We then applied CSI-Microbes to analyze each dataset separately. First, we analyzed the breast cancer samples. Five of the 11 patients from (Chung et al. 2017) and four of the six patients from (Karaayvaz et al. 2018) contained a sufficient number (≥ 5) of tumor and non-tumor cells to analyze using CSI-Microbes. From these nine samples, CSI-Microbes identified the genus *Streptomyces* as differentially abundant between the tumor and non-tumor cells in one sample, BC06 (FDR corrected p-value =.008) (Figure 2A). The non-tumor cells in BC06 consist of stromal and B cells, and the differential abundance of *Streptomyces* is driven by the difference between the tumor cells and the B cells. At the species level, CSI-Microbes identified two *Streptomyces* species to be higher in the tumor cells compared to the B cells (*Streptomyces novaecaesareae*, FDR=.023 and *Streptomyces specialis*, FDR=.035) (**Figure 2A**). To validate this result, we leveraged the capability of CSI-Microbes to use the RNA spike-in data provided in the (Chung et al. 2017) dataset. Across all samples in this cohort, the abundance of the *Streptomyces* genus is positively correlated with the abundance of RNA spike-in transcripts, indicating that in these samples this genus is a contaminant (Spearman-rank correlation, rho=.17, FDR=.007) (Methods). However, there is no correlation between *Streptomyces* abundance and RNA spike-in transcript levels across cells from patient BC06 (rho=-.63, FDR=.22). A recent survey of the tumor microbiome across TCGA found reads belonging to the *Streptomyces* genus in all but one of the breast cancer samples and high abundance of Streptomyces in ~10% (116/1086) of all breast cancer samples (Poore et al., 2020). Together with its cell type specificity in that patient, these findings testify that the presence of *Streptomyces* is not a result of sample contamination and that it is located intracellularly within the tumor cells.

In the previous datasets, the cells were usually unsorted so individual plates generally contained both tumor and non-tumor cells, which CSI-Microbes uses to account for plate-specific contaminants. In the melanoma cohort we analyzed, the cells were sorted prior to sequencing, which resulted in cancerous and non-cancerous cells being sequenced separately. To minimize the effects of potential contamination, we analyzed only the eight melanoma patients whose tumor cells were sequenced in a minimum of two plates, which was sufficient to control for contamination in the validation dataset (see Methods). In these eight melanoma patients, CSI-Microbes detected eight differentially abundant microbes at the genus level in two of the patients. In sample Mel129pa, the bacterial genus *Propionibacterium* was more abundant in tumor cells than in non-tumor cells (FDR=.04) (Figure 2B). *Propionibacterium* is a human skin commensal that has been reported to be an intracellular pathogen in multiple diseases, including prostate cancer (McDowell et al., 2013). 103 of the 469 melanoma samples in TCGA have reads belonging to the *Propionibacterium* genus, which is highly abundant in four of these samples (Poore et al., 2020). In patient Mel106, CSI-Microbes identified six differentially abundant microbes at the genus level: four fungi (*Coprinopsis*, *Heterobasidion*, *Phanerochaete* and

Laccaria) and two bacteria (*Chryseobacterium* and *Brevundimonas*). In each of the four fungi genera, CSI-Microbes identified one differentially abundant fungus species (*Coprinopsis cinereal*, *Heterobasidion irregulare*, *Phanerochaete carnosa*, *Laccaria bicolor*). Interestingly, it has been reported that most pathogenic fungi have an intracellular phase (Gladieux et al., 2017). *Coprinopsis cinereal*, the species with the strongest enrichment (FDR < 3e-8), has been reported as an uncommon human pathogen (Correa-Martinez et al., 2018) (**Figure 2B**).

Finally, we analyzed the 12 patients with sufficient tumor and non-tumor cells from the two brain cancer datasets. Notably, CSI-Microbes did not find any differentially abundant microbes across these 12 patients.

Figure 2

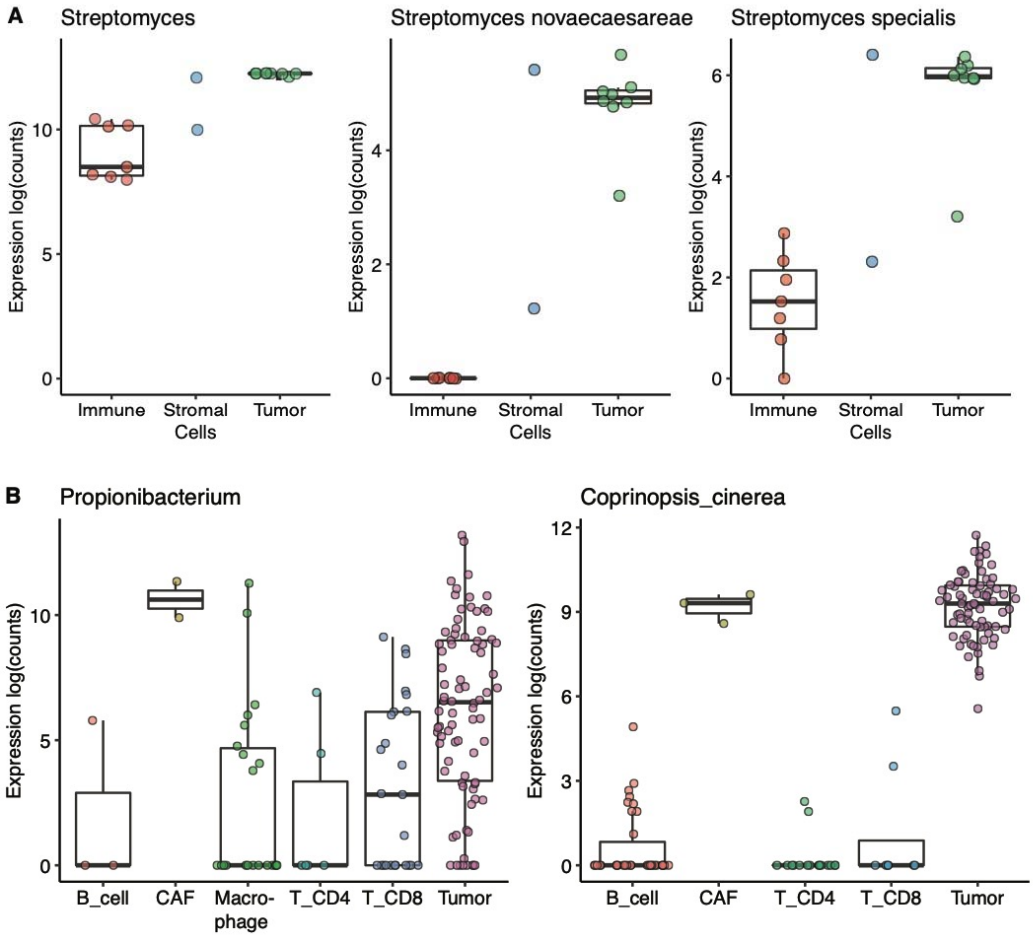

Figure 2. **(A)** Comparison of the abundance of the *Streptomyces* genus and the *Streptomyces novacaesareae* and *specialis* species between immune, stromal and tumor cells in breast cancer patient BC06 **(B)** Comparison of the abundance of the *Propionibacterium* genus between cell types in sample Mel129pa and the abundance of *Coprinopsis cinereal* in patient Mel106 in the melanoma dataset.

Discussion

To the best of our knowledge, this paper introduces the first approach for identifying cell type-specific intracellular bacteria from scRNA-seq data. We first demonstrate that our approach can distinguish cells infected with intracellular bacteria from uninfected cells. Next, we apply our approach to scRNA-seq datasets from multiple cancer types, which yields several novel bioinformatic based discoveries of putative intracellular pathogens whose abundance is associated with specific cell types in different tumor types, varying between different samples within a cohort of a similar cancer type. We find that species belonging to the *Streptomyces* genus are specifically associated with tumor cells in a breast cancer tumor. Finally, we find bioinformatic evidence suggesting the existence of several intracellular bacteria and fungi in melanoma tumor cells. Our analysis of brain tumors did not identify any differentially abundant microbial genera.

This paper represents a proof-of-concept for a new computational pipeline for identifying cell type-specific intracellular bacteria. To establish this approach, we analyzed previously generated datasets that did not explicitly control for contamination. Although CSI-Microbes was carefully designed to control for contamination and provides multiple validation approaches, many of the previously published datasets regrettably lack some of these validation measures, at least in some of the samples. In addition, the sequencing protocols used to generate the datasets are using polyA mRNA selection, which enriches for eukaryotic mRNA. As only some prokaryotic mRNA transcripts are polyadenylated, the microbial transcripts identified by CSI-Microbes thus may represent an under-sampling of the likely intracellular microbial diversity of human tumors.

The computational findings presented here call for further experimental testing and validation, e.g., using RNAscope (Wang et al. 2012) to actually visualize the presence of the microbes in the tumor cells. If some of our computational predictions are indeed successfully validated, then an important and open question would follow, 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? The identification of genera or species that are highly specific inhabitants of tumor cells is of particular interest in this respect.

Methods

The inputs to CSI-Microbes are FASTQ files from scRNA sequencing and sample metadata including cell type annotations and any known sources of contamination such as the sequencing plate. CSI-Microbes outputs a ranked list of differentially abundant microbes and their discriminatory power between cell types using the following steps: **(1a)** The input FASTQ file (if using unpaired reads) or files (if using paired reads) are converted to a BAM file. If the empty wells or spike-in validations are used, the FASTQ files are converted to a BAM file aligned to the human genome (hg38) and any relevant spike-in sequences using STAR two-pass mode (v2.7.3.a) (Dobin et al., 2013). Otherwise, the FASTQ files are converted to BAM files using the FastqToSam command from picard (v2.22.2). **(1b)** The BAM file is filtered for human reads, aligned to a large number of bacteria, viral, archaea, fungi and protozoan genomes and scored using GATK PathSeq (v4.1.6.0) and the GATK PathSeq resource bundle (retrieved 1/21/2019)

(Walker et al., 2018). **(2)** The unambiguous read counts at the specified taxonomic level are extracted to form an OTU-by-cell integer matrix. Cell types with fewer than five cells are removed and operational taxonomic units (OTUs) with fewer than three reads in the minimum number of cells are removed. The minimum number of cells is 70% of the number of cells in the least frequent cell type. The reads are normalized for sequencing depth by pooling cells belonging to the same cell type using the *computeSumFactors* function from the R package scran (v1.14.1) (Lun et al., 2016). **(3)** OTUs without log count greater than 2 in the minimum number of cells are removed to minimize FDR penalty. Differentially abundant microbes are identified using the two-sided Wilcoxon rank sum test with a log-fold-change greater than one while controlling for the plate as implemented in the findMarkers function in the R package scran (Lun et al., 2016). **(4a)** If RNA spike-in transcripts were used in the experiment, CSI-Microbes uses their abundance level to validate differentially abundant microbes using the principle that the abundance levels of a contaminant genus or species should correlate with the RNA spike-in levels across wells in a given plate (Davis et al., 2018). Following this principle, we check whether there is a statistically significant correlation between the abundance of the RNA spike-in transcripts and any identified differentially abundant OTU using Spearman rank-correlation and FDR < .05. Analogously, we repeat the differential abundance test using the RNA spike-in transcript normalization using the *computeSpikeFactors* function from scran. **(4b)** If empty wells were sequenced as part of the experiment, CSI-Microbes will check whether differentially abundant OTUs are present in the empty wells at high frequency.

CSI-Microbes controls for known sources of contamination, such as the sequencing plate, by computing the Wilcoxon rank-sum test between cell types within a sequencing plate and combining the p-values using Stouffer's weighted z-score method. However, several large scRNA-seq cancer studies first sort cells using cell-surface markers and FACS before sequencing cells, which results in cell types being sequenced in different plates, as occurring in the melanoma dataset analyzed in this paper. In such a situation, it is not possible compute the Wilcoxon rank-sum test between cell types within a sequencing plate. In these situations, CSI-Microbes can still be used if the cell type of interest has been sequenced across at least two plates; we assume that the same contamination is unlikely to affect both plates without affecting all plates in the study. In this situation, CSI-Microbes using the one-sided Wilcoxon rank sum test is run once for each pair of tumor-non-tumor plates, and the resulting FDR-corrected p-values are combined by taking the maximum value. We simulated such a situation using the gold-standard *Salmonella* dataset, which was sequenced across four plates, and found that this approach identified *Salmonella* as the only differentially abundant microbe between the cell types. We employed this correction procedure only in the melanoma analysis.

Archaea, bacterial and viral read counts from TCGA were downloaded from (Poore et al., 2020). A genus is defined to be *highly abundant* in a sample if the normalized abundance in the sample is greater than the median abundance across all tumor types plus three median absolute deviations (MADs). If x equals the number of reads belonging to the genus g, then g is highly abundant if $\log2(x+1) > \text{median} + 3 * \text{MADs}$.

Acknowledgements
This research was supported by the Intramural Research Program of the National Institutes of Health, National Cancer Institute. This work utilized the computational resources of the NIH

HPC Biowulf cluster. (http://hpc.nih.gov). The authors would like to thank Wolfgang Resch for assistance running this analysis on the NIH HPC Biowulf cluster. The authors would like to thank Leigh Greathouse, Hector Corrado-Bravo, Norma Andrews, Steve Christensen, Max Leiserson, Itay Tirosh, Livnat Jerby-Arnon, Frank Lowery and Sri Krishna for useful discussions about this project. WR's contribution to this research was supported in part by NSF award DGE-1632976.

References

Aulicino, A., Rue-Albrecht, K. C., Preciado-Llanes, L., Napolitani, G., Ashley, N., Cribbs, A., Koth, J., Lagerholm, B. C., Ambrose, T., Gordon, M. A., Sims, D., & Simmons, A. (2018). Invasive *Salmonella* exploits divergent immune evasion strategies in infected and bystander dendritic cell subsets. *Nature Communications*, *9*(1), 4883. https://doi.org/10.1038/s41467-018-07329-0

Bossel Ben-Moshe, N., Hen-Avivi, S., Levitin, N., Yehezkel, D., Oosting, M., Joosten, L. A. B., Netea, M. G., & Avraham, R. (2019). Predicting bacterial infection outcomes using single cell RNA-sequencing analysis of human immune cells. *Nature Communications*, *10*(1), 3266. https://doi.org/10.1038/s41467-019-11257-y

Bullman, S., Pedamallu, C. S., Sicinska, E., Clancy, T. E., Zhang, X., Cai, D., Neuberg, D., Huang, K., Guevara, F., Nelson, T., Chipashvili, O., Hagan, T., Walker, M., Ramachandran, A., Diosdado, B., Serna, G., Mulet, N., Landolfi, S., Ramon, S., … Meyerson, M. (2017). Analysis of *Fusobacterium* persistence and antibiotic response in colorectal cancer. *Science*, *358*(6369), 1443–1448. https://doi.org/10.1126/science.aal5240

Castellarin M., Warren, R. L., Freeman, J. D., , Dreolini, L., Krzywinski, M., Strauss, J., Barnes, R., Watson, P., Allen-Vercoe, E., Moore, R. A., & Holt, R. A. (2012). *Fusobacterium nucleatum* infection is prevalent in human colorectal carcinoma. *Genome Research*, *22*(2), 299–306. https://doi.org/10.1101/gr.126516.111.Freely

Chung, W., Eum, H. H., Lee, H. O., Lee, K. M., Lee, H. B., Kim, K. T., Ryu, H. S., Kim, S., Lee, J. E., Park, Y. H., Kan, Z., Han, W., & Park, W. Y. (2017). Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nature Communications*, 8(5), 15081. https://doi.org/10.1038/ncomms15081

Correa-Martinez, C., Brentrup, A., Hess, K., Becker, K., Groll, A. H., & Schaumburg, F. (2018). First description of a local *Coprinopsis cinerea* skin and soft tissue infection. *New Microbes and New Infections*, *21*, 102–104. https://doi.org/10.1016/j.nmni.2017.11.008

Darmanis, S., Sloan, S. A., Croote, D., Mignardi, M., Chernikova, S., Samghababi, P., Zhang, Y., Neff, N., Kowarsky, M., Caneda, C., Li, G., Chang, S. D., Connolly, I. D., Li, Y., Barres, B. A., Gephart, M. H., & Quake, S. R. (2017). Single-cell RNA-seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. *Cell Reports*, *21*(5), 1399–1410. https://doi.org/10.1016/j.celrep.2017.10.030

Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., & Callahan, B. J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and

metagenomics data. *Microbiome*, *6*(1), 1–14. https://doi.org/10.1186/s40168-018-0605-2

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Geller, L. T., Barzily-rokni, M., Danino, T., Jonas, O. H., Shental, N., Nejman, D., Gavert, N., Zwang, Y., Cooper, Z. A., Shee, K., Thaiss, C. A., Reuben, A., Livny, J., Avraham, R., Frederick, D. T., Ligorio, M., Chatman, K., Johnston, S. E., Mosher, C. M., … Straussman, R. (2017). Potential role of intratumor bacteria in mediating tumor resistance to the chemotherapeutic drug gemcitabine. *Science*, 357(6356), 1156–1160.

Gladieux, P., Byrnes, E. J., Aguileta, G., Fisher, M., Billmyre, R. B., Heitman, J., & Giraud, T. (2017). Epidemiology and evolution of fungal pathogens in plants and animals. In *Genetics and Evolution of Infectious Diseases: Second Edition* (pp. 71–98). https://doi.org/10.1016/B978-0-12-799942-5.00004-4

Gur, C., Ibrahim, Y., Isaacson, B., Yamin, R., Abed, J., Gamliel, M., Enk, J., Bar-On, Y., Stanietsky-Kaynan, N., Coppenhagen-Glazer, S., Shussman, N., Almogy, G., Cuapio, A., Hofer, E., Mevorach, D., Tabib, A., Ortenberg, R., Markel, G., Miklić, K., … Mandelboim, O. (2015). Binding of the Fap2 protein of *Fusobacterium nucleatum* to human inhibitory receptor TIGIT protects tumors from immune cell attack. *Immunity*, *42*(2), 344–355. https://doi.org/10.1016/j.immuni.2015.01.010

Gur, C., Maalouf, N., Shhadeh, A., Berhani, O., Singer, B. B., Bachrach, G., & Mandelboim, O. (2019). *Fusobacterium nucleatum* supresses anti-tumor immunity by activating CEACAM1. *OncoImmunology*, *8*(6), e1581531. https://doi.org/10.1080/2162402X.2019.1581531

Jerby-Arnon, L., Shah, P., Cuoco, M. S., Rodman, C., Su, M. J., Melms, J. C., Leeson, R., Kanodia, A., Mei, S., Lin, J. R., Wang, S., Rabasha, B., Liu, D., Zhang, G., Margolais, C., Ashenberg, O., Ott, P. A., Buchbinder, E. I., Haq, R., … Regev, A. (2018). A cancer cell program promotes T cell exclusion and resistance to checkpoint blockade. *Cell*, *175*(4), 984-997.e24. https://doi.org/10.1016/j.cell.2018.09.006

Kapadia, M., Rolston, K. V. I., & Han, X. Y. (2007). Invasive *Streptomyces* infections: Six cases and literature review. *American Journal of Clinical Pathology*, *127*(4), 619–624. https://doi.org/10.1309/QJEBXP0BCGR54L15

Karaayvaz, M., Cristea, S., Gillespie, S. M., Patel, A. P., Mylvaganam, R., Luo, C. C., Specht, M. C., Bernstein, B. E., Michor, F., & Ellisen, L. W. (2018). Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. *Nature Communications*, *9*(1), 3588. https://doi.org/10.1038/s41467-018-06052-0

Kostic, A. D., Gevers, D., Pedamallu, C. S., Michaud, M., Duke, F., Earl, A. M., Ojesina, A. I., Jung, J., Bass, A. J., Tabernero, J., Baselga, J., Liu, C., Shivdasani, R. A., Ogino, S., Birren, B. W., Huttenhower, C., Garrett, W. S., & Meyerson, M. (2012). Genomic analysis identifies association of *Fusobacterium* with colorectal carcinoma. *Genome Research*, *22*(2), 292–298. https://doi.org/10.1101/gr.126573.111

Lun, A. T. L., Bach, K., & Marioni, J. C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biology*, *17*(1), 1–14. https://doi.org/10.1186/s13059-016-0947-7

McDowell, A., Patrick, S., Eishi, Y., Lambert, P., & Eady, A. (2013). *Propionibacterium acnes* in human health and disease. *BioMed Research International*, *2013*, 10–12. https://doi.org/10.1155/2013/493564

Pleguezuelos-Manzano, C., Puschhof, J., Rosendahl Huber, A., van Hoeck, A., Wood, H. M.,

Nomburg, J., Gurjao, C., Manders, F., Dalmasso, G., Stege, P. B., Paganelli, F. L., Geurts, M. H., Beumer, J., Mizutani, T., Miao, Y., van der Linden, R., van der Elst, S., Ambrose, J. C., Arumugam, P., … Clevers, H. (2020). Mutational signature in colorectal cancer caused by genotoxic *pks* $^{+}$ *E. coli*. *Nature*, *580*(7802), 269–273. https://doi.org/10.1038/s41586-020-2080-8

Poore, G. D., Kopylova, E., Zhu, Q., Carpenter, C., Fraraccio, S., Wandro, S., Kosciolek, T., Janssen, S., Metcalf, J., Song, S. J., Kanbar, J., Miller-Montgomery, S., Heaton, R., Mckay, R., Patel, S. P., Swafford, A. D., & Knight, R. (2020). Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature*, *579*(7800), 567–574. https://doi.org/10.1038/s41586-020-2095-1

See, P., Lum, J., Chen, J., & Ginhoux, F. (2018). A single-cell sequencing guide for immunologists. *Frontiers in Immunology*, *9*, 2425. https://doi.org/10.3389/fimmu.2018.02425

Venteicher, A. S., Tirosh, I., Hebert, C., Yizhak, K., Neftel, C., Filbin, M. G., Hovestadt, V., Escalante, L. E., Shaw, M. L., Rodman, C., Gillespie, S. M., Dionne, D., Luo, C. C., Ravichandran, H., Mylvaganam, R., Mount, C., Onozato, M. L., Nahed, B. V., Wakimoto, H., … Suvà, M. L. (2017). Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science*, *355*(6332), pii: eaai8478. https://doi.org/10.1126/science.aai8478

Walker, M. A., Pedamallu, C. S., Ojesina, A. I., Bullman, S., Sharpe, T., Whelan, C. W., & Meyerson, M. (2018). GATK PathSeq: a customizable computational tool for the discovery and identification of microbial sequences in libraries from eukaryotic hosts. *Bioinformatics*, *34*(24), 4287–4289. https://doi.org/10.1093/bioinformatics/bty501

Wang, F., Flanagan, J., Su, N., Wang, L. C., Bui, S., Nielson, A., Wu, X., Vo. H. T., Ma, X. J., Luo, Y. (2012). RNAscope: a novel in situ RNA analysis platform for formalin-fixed, paraffin-embedded tisues. *Journal of Molecular Diagnostics* , *14*(1), 22-29. doi: 10.1016/j.jmoldx.2011.08.002.

Yu, T. C., Guo, F., Yu, Y., Sun, T., Ma, D., Han, J., Qian, Y., Kryczek, I., Sun, D., Nagarsheth, N., Chen, Y., Chen, H., Hong, J., Zou, W., & Fang, J. Y. (2017). *Fusobacterium nucleatum* promotes chemoresistance to colorectal cancer by modulating autophagy. *Cell*, *170*(3), 548-563.e16. https://doi.org/10.1016/j.cell.2017.07.008