ABSTRACT

Title of Dissertation:	COMPUTATIONAL METHODS FOR THE IDENTIFICATION OF MUTATION SIGNATURES AND INTRACELLULAR MICROBES IN CANCER
	Wells Ivens Robinson, Doctor of Philosophy, 2021
Dissertation directed by:	Professor Max Leiserson, Department of Computer Science Professor Eytan Ruppin, Department of Computer Science

Cancer is the second leading cause of death in the United States behind heart disease, killing ~600,000 Americans per year. Technological advances have lowered the cost of sequencing a tumor genome even faster than would have been predicted by Moore's law. However, specialized computational techniques are required to effectively analyze this genomic data. In this dissertation, we present two such computational approaches to address key challenges in the field of computational cancer biology. Given the importance of reproducibility in biomedical research, we provide publicly available workflows for reproducing the results from our computational approaches.

In the first part of this thesis, we present a novel approach for the extraction of mutation signatures from matrices of somatic mutations. One computational

challenge for extracting mutation signatures is the relatively small number of mutations in each tumor compared to the relatively large number of distinct signatures, which can be mathematically similar to each other. To help address this computational challenge, we apply ideas from the field of topic modeling to develop the first mutation signature model, the Tumor Covariate Signature Model (TCSM), that can incorporate known tumor covariates. We focus on two mathematically similar signatures associated with distinct covariates to evaluate TCSM and show that by leveraging these covariates, TCSM can more accurately distinguish between mutations attributed to these two signatures than existing NMF-based approaches.

The second part focuses on the microbes in the tumor microenvironment. It is not currently known whether microbial reads identified from tumor sequencing datasets result from contamination or represent either extracellular or intracellular microbial residents of the tumor microenvironment. We develop a computational approach named **CSI-Microbes** (computational identification of **C**ell type **S**pecific **I**ntracellular **Microbes**) that mines single-cell RNA sequencing (scRNA-seq) datasets to distinguish cell-type specific intracellular microbes from other microbes. We show that CSI-Microbes can identify previously reported intracellular microbes from both human-designed and cancer scRNA-seq datasets. Finally, we apply CSI-Microbes to a large scRNA-seq lung cancer dataset and identify microbial taxa in tumor cells with a transcriptomic signature of decreased immune activity.

COMPUTATIONAL METHODS FOR THE IDENTIFICATION OF MUTATION SIGNATURES AND INTRACELLULAR MICROBES IN CANCER

by

Wells Ivens Robinson

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2021

Advisory Committee:
Professor Max Leiserson, Chair
Professor Eytan Ruppin, Co-Chair
Professor Rob Patro
Professor Furong Huang
Professor Najib El-Sayed

Acknowledgements

First and foremost, I'd like to thank my co-advisors Eytan Ruppin and Max Leiserson. Throughout an ever-changing PhD, having Eytan as my advisor was the one constant. He gave me a tremendous amount of freedom to learn and develop into the scientist that I am today while always being available to chat and lead me back in the right direction when I strayed too far. Max is the perfect complement to Eytan. He guided me masterfully through the field of mutation signatures and my first-ever first-author journal publication. I've learned so much from both of them and consider myself very fortunate to have two mentors who care deeply about me as a scientist and a person.

Next, I would like to thank the members of the Ruppin Lab and Center for Bioinformatics and Computational Biology (CBCB) who mentored me at the beginning of my PhD. In particular, I appreciate both the advice and friendship provided by Joo Sang Lee, Noam Auslander, Erez Persi and Justin Wagner during this time. I would like to thank all of my professors at the University of Maryland and in particular, Amol Deshpande, David Van Horn, Jim Reggia and Norma Andrews for being very generous with their time at the beginning of my Ph.D. I would like to thank all members of both the Cancer Data Science Lab (formerly the Ruppin Lab) and the Leiserson Research Group for being such wonderful colleagues and friends. In particular, I'd like to thank Rotem Katzir, Nishanth Nair, Kun Wang, Kuoyuan Cheng, Sanju Sinha, Sushant Patkar, David Crawford, Mike Gertz, Jason Fan and Mark Keller. I am particularly grateful to Alejandro Shaffer for supporting and

mentoring me during the end of my Ph.D., particularly in guiding me through countless low-level details for CSI-Microbes. I am also grateful to Fiorella Schischlik for being both a great friend and great co-author on CSI-Microbes. I'd also like to thank my collaborators in the Surgery Branch including Nick Restifo, Stephen Rosenberg, Suman Vodnala, Paul Robbins, Jared Gartner, Sri Krishna and especially Frank Lowery for teaching me so much about cancer immunotherapy and allowing me to participate in clinical rounds. Meeting patients battling cancer has been one of the most humbling and motivating experiences of my life.

Finally, I would like to thank my friends and family for supporting me through the last six years. It has been extremely valuable to have such a wonderful community to constantly remind me that there is life beyond science. In particular, I'd like to thank my girlfriend Hannah for putting up with me spending way too much of our date nights discussing "cells in wells". And of course, I'd like to thank both my mom, my dad and my sister for being such a wonderful constant in my life since I was born.

Table of Contents

List of Figures

Chapter 1: Introduction

1.1 Background

The rise of next-generation sequencing (NGS) has revolutionized the entire field of biology including cancer genomics[1,2]. NGS is massively parallel, high-throughput DNA and RNA sequencing. The advent of NGS is largely responsible for the dramatic decrease from ~$2.7 billion and ~10 years to sequence the first human genome in 2001 to ~$1,000 and a few days to sequence a genome today, which significantly outpaces even Moore's law[3].

At a high level, NGS starts with either DNA or RNA as input material and outputs thousands to millions of "short reads" where each read is a string of ~50-250 characters from an alphabet of four characters, each of which represent one type of nucleotide (adenine ("A"), thymine ("T"), guanine ("G") and cytosine ("C")). Next, these short sequences are either "assembled" into a set of long contiguous regions (usually representing a genome)[4] or "aligned" to an existing reference genome[5–7]. Read alignment is generally solved using variants on string matching algorithms from computer science[8]. In string matching terms, the read alignment problem is the identification of exact or near-exact occurrences of pattern *R* within text *G* where *R* represents the short DNA or RNA read and *G* represents the large genome.

NGS applied to DNA is called DNA-sequencing and at a high level, can be partitioned into targeted sequencing, which is most commonly targeted to the exome, which is the ~2% of the genome that encodes proteins (whole exome sequencing or WXS), and untargeted or whole genome sequencing (WGS)[9]. The primary objective

of DNA-sequencing is to identify mutations, which are differences between the genome or exome being sequenced and a reference genome, using mutation calling algorithms such as VarScan 2[10,11] and MuTect2[12]. Mutations can broadly be divided into germline, which are passed down from parent to child, and somatic, which occur during the lifetime of an individual in a single cell and are only passed down to daughter cells. In cancer genomics, the objective is to identify somatic mutations that occur in the cancerous cells.

NGS applied to RNA is called RNA-sequencing and is most commonly targeted to the messenger RNA (mRNA) molecules, which represent the intermediate stage between DNA and protein[13]. The primary objective of RNA-sequencing is to quantify the number of transcripts being transcribed from a given gene using algorithms that align reads to the transcriptome[6]. One challenge for analyzing DNA and RNA-sequencing from tumor samples is that tumor samples contain a mix of tumor and non-tumor cells and reads from both of these cell populations are intermixed in the DNA and RNA-sequencing output[14]. To avoid this problem of mixing multiple cell-types, techniques have been developed to sequence the RNA[15] or the DNA[16] of a single cell. Very recent technological advances have scaled the number of single cells able to be sequenced in a single experiment from one in 2009 to tens of thousands today[17].

By time that I started my PhD in 2016, algorithmic development paired with the application of NGS to tens of thousands of cancer genomes by international consortiums such as The Cancer Genome Atlas (TCGA)[18–20] and the International Cancer Genome Consortium (ICGC)[21] had already yielded many novel discoveries

about the drivers and hallmarks of cancer[14]. For example, large sequencing datasets has enabled the identification of recurrently mutated genes in specific cancer types in unexpected pathways like splicing and protein homeostasis[22,23]. One unexpected finding is that while there are only a small number of genes that are frequently mutated across many tumors, there are a large number of genes that are infrequently mutated[14].

Some of these somatic mutations drive the cancer by either transforming one class of proteins (oncogenes) into hyperactive versions of themselves that cause the cell to grow uncontrollably or by transforming another type of proteins (tumor suppressors) into non-functional versions of themselves, which are no longer able to stop the cell from growing uncontrollably[14]. However, these somatic mutations can also be recognized as foreign by the immune system, which has been exploited by the field of cancer immunotherapy that I will mention later in this introduction. One prominent computational research area has been the development of computational tools to distinguish the small number of somatic mutations that "drive" the cancer ("driver mutations") from the many somatic mutations that do not play a functional role in the development or progression of cancer ("passenger mutations"). One group of tools looks for single genes that are more mutated than expected given the background rate of mutation[24,25]. Another set of computational tools like CoMEt[26] and HotNet2[27] looks for sets of driver genes using mutual exclusivity or network propagation approaches. While these "passenger mutations" do not drive the cancer, they provide a functional readout of the mutational processes active in the tumor. I

became interested in mutational processes after taking Max Leiserson's class on Machine Learning for Cancer Mutations in the fall of 2017.

1.2 Mutational Processes in Cancer

Somatic mutations are caused by either mutagenic processes such as ultraviolet radiation and smoking or defective DNA repair processes such as mismatch repair and homologous recombination[28]. Generally, the somatic mutations in a tumor are thought to occur from multiple co-occurring mutagenic processes[29]. For example, an inactivating somatic mutation in a gene in a DNA repair pathway can inactivate that pathway, which can cause additional mutations. The inactivation of the homologous recombination repair (HR) pathway is one of the most clinically relevant DNA repair pathway defects because tumors with defects in this pathway are particularly susceptible to treatment with PARP inhibitors[30]. The biallelic inactivation of either *BRCA1* or *BRCA2* is one of the most common defects to the HR pathway although defective HR pathway has been reported in the absence of these inactivation[31].

It has long been recognized that some of these mutagenic processes cause specific patterns of mutations such as the dramatic increase in the number of G to T substitutions in lung tumors of smokers compared to those of non-smokers[32]. Incredibly, Alexandrov *et al.*[33] showed that the application of unsupervised machine learning approaches to thousands of cancer exomes and genomes sequenced by TCGA could extract mathematical patterns of mutations termed "mutation signatures", some of which strongly resemble the previously reported patterns of

known mutagenic processes. In their approach, the catalog of single-base substitutions, which are a subset of all somatic mutations, from a tumor are first categorized into 96 categories, where a category is defined by the base substitute and the immediately flanking nucleotides (ex. C[G>T]C represents a T substituted for a G flanked on either side by a C). Next, each set of 96 categories from patients are concatenated together into a mutation count matrix of size N-by-96 where N equals the number of patients. Finally, this mutation count matrix is deconvolved using the machine learning approach non-negative matrix factorization (NMF) into two smaller matrices, one matrix representing the identified mutation signatures and one matrix representing each patient's exposure to each mutation signature (**Figure 1**).

Shiraishi *et al.*[34] first observed that this problem could also be solved using machine learning approaches from the field topic modeling, which tries to identify topics present across a large cohort of text documents. In this analogy, the cancer genomes are documents, mutation signatures are topics, exposures are topic prevalence and mutation categories are words. The field of topic modeling has generated a significant number of new models and extensions to the original latent Dirichlet allocation (LDA)[35] including correlated topic models[36] as well as supervised[37] and semi-supervised[38] models. We and others[34,39] have utilized the connection between the two fields to apply models originally developed for topic modeling to mutation signatures.

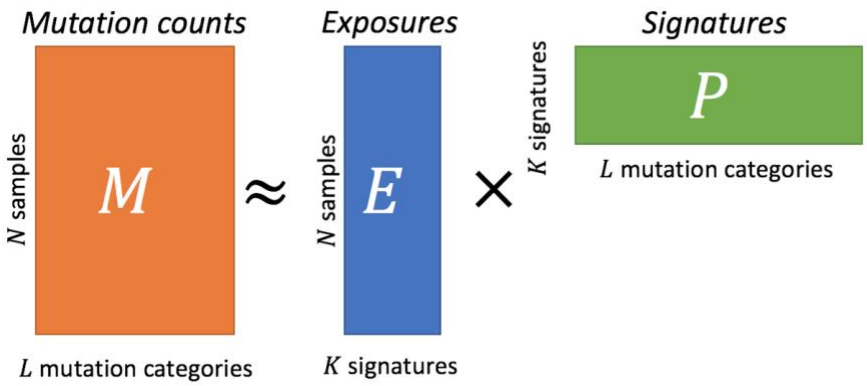


Figure 1: Overview of mutation signature extraction
M, the mutation count matrix (orange) is deconvolved into two smaller matrices: P, which represents the mutation signatures (green) and E, which represents the exposures of each mutation signature in each of the N samples (blue).

1.3 Tumor Covariate Signature Model

Research goal: develop a mutation signature model that incorporates tumor covariates and better distinguishes between mathematically similar mutation signatures.

We were particularly interested in applying semi-supervised topic modeling approaches to the problem of mutation signature extracting because we noticed that the existing unsupervised NMF-based approaches struggled to distinguish between mutations caused by mathematically similar mutation signatures. This problem was particularly pronounced when using the much smaller number of mutations called from whole exome sequencing (WXS), where only ~2% (the protein coding region) of the genome is sequenced[9]. We were particularly interested in this problem because signature 3, which has been proposed as a biomarker for PARP inhibitors because it is associated with defects in the aforementioned HR pathway, is mathematically very similar to signature 5[40]. To address this problem, we introduce the Tumor Covariate Signature Model (TCSM), which is the first approach to mutation signatures that can

incorporate known tumor metadata and demonstrate the improved performance of our approach in distinguishing between similar signatures on simulated and real data[41].

1.4 Transition to Cancer Immunotherapy

During my PhD, my co-advisor, Eytan Ruppin, moved from the University of Maryland to the National Cancer Institute to start the Cancer Data Science Laboratory (CDSL). As a member of the CDSL, I began a small collaboration with the lab of Nick Restifo in the Surgery Branch. The Surgery Branch is one of the pioneers in the development of cancer immunotherapy, which uses the patient's own immune system to attack their cancer[42].

During this collaboration, Nick invited me to attend the Surgery Branch's clinical rounds where I met several cancer patients who were being treated by one of their clinical trials. After this experience, when Nick offered me a position in his lab, I immediately accepted even though it meant moving to a completely new area of research and leaving the field of mutation signatures without completing the multiple ideas and extensions that Max and I had planned to do after the publication of TCSM.

Although Nick Restifo unfortunately left the Surgery Branch only a few months after I joined, I was already committed to the work being done there and continued to attend both clinical rounds and their lab meetings and journal clubs. It was during one of these journal clubs when I became interested in the tumor microbiome.

1.5 The Tumor Microbiome

The human microbiome is the collection of microbes, which includes bacteria and viruses, throughout the human body[43]. It has been estimated that the human body contains roughly the same number of microbial cells as human cells[44]. Early studies of the human microbiome relied on the study of microbes that were cultured from human tissue and waste[45]. However, this culturing-based approach is limited because it can only be used to study microbes capable of growing in the provided culture media[45]. In contrast, NGS is able to provide an unbiased sampling of the genomic sequences of the microbes present in a sample[43]. Very recently, interest has grown in the tumor microbiome, which is the collection of microbes present in the tumor microenvironment[46]. Several recent papers have pointed to the functional importance of the tumor microbiome in both progression and response to treatment of tumors[47–49]. The tumor microbiome has previously been studied computationally from NGS using a "computational transcriptome subtraction" approach where sequencing reads are first aligned to the human reference genome and unmapped, high-quality reads are subsequently aligned against a large database of many microbial genomes[50,51]. This approach led to the landmark discovery that the clonal integration of a previously unknown polyomavirus (*Merkel polyomavirus*) causes ~80% of Merkel cell carcinomas, which is a rare but aggressive human skin cancer[52]. Similar computational approaches have been applied to identify the enrichment of the bacterial genus *Fusobacterium* in colorectal carcinoma compared to matched normal tissues[48,53].

The journal club paper that piqued my interest in the tumor microbiome followed up on these papers and showed that *Fusobacterium* can exist intracellularly within colorectal carcinoma cells[47]. The existence of *Fusobacterium* within these tumor cells meant that peptides from this bacterium, similar to peptides derived from mutated proteins, should be presented by the tumor cells and recognized as foreign by the immune system. To both my collaborators in the Surgery Branch and myself, this meant that we could target this bacterium (using immune cells) and treat patients with colorectal carcinomas.

From a computational perspective, moving from mutation signatures to the tumor microbiome meant working with significantly larger datasets. Although the mutation count matrix input to mutation signature extraction is derived from NGS data, the identification of somatic mutations from NGS data is usually the primary outcome of sequencing, and mutation calling is always performed by the original authors. The main pre-processing step for mutation signature extraction is the conversion of the mutation calls from VCF files (average VCF file size ~ 10 megabytes) to the mutation count matrix. The development of a Snakemake pipeline to apply the same standardized approach to tens of datasets with Max and a very talented undergraduate student Mark Keller was one of my contributions to the Leiserson Research Group. In contrast, the sequencing of microbial reads is usually an accidental by-product of sequencing to identify mutations or characterize the transcriptome and these microbial reads are usually ignored or filtered out by the original authors. Therefore, the analysis of microbial reads from NGS datasets usually

needs to start with the raw reads in FASTQ format (average FASTQ file size ~10 gigabytes), which are ~3 orders of magnitude larger.

My first project in this field was the application of an existing computational host subtraction approach[54] to hundreds of tumor samples from patients treated by the Surgery Branch to identify the subset of tumors with reads that mapped to *Fusobacterium*. I completed this analysis and provided the list of tumors to my collaborators in the Surgery Branch, who plan to look for immune cells that recognize *Fusobacterium* in these tumors. At the same time, Eytan pushed me to look for additional intracellular bacteria in these tumor samples. However, no definitive list of intracellular microbes existed at the time because research is biased towards a small number of disease-causing bacteria, which excludes most members of the human microbiome, and the distinction between obligate intracellular bacteria, which can only reproduce inside of a host cell, and facultative intracellular bacteria, which can reproduce both inside and outside of a host cell is blurry (personal correspondence with Norma Andrews). Given this background, we began to brainstorm a computational approach for the identification of intracellular bacteria.

1.6 CSI-Microbes

Research goal: develop a computational model to distinguish intracellular microbes from extracellular and contaminating microbes.

To identify intracellular microbes from NGS, it is necessary to distinguish them from both extracellular microbes, which are microbes living outside of host cells, and contaminating microbes, which are microbes not originally present in the tissue that are introduced prior to sequencing[55,56]. Previous computational approaches

to remove contaminating microbes rely on the idea that samples that are processed together should have similar contaminates[57,58]. However, such approaches usually rely on one or at most two NGS samples per patient. In contrast, CSI-Microbes uses single cell RNA-sequencing, which sequences hundreds to thousands of cells from multiple cell-types per patient. We compare the (normalized) number of microbial reads between cells from different cell-types under the assumption that the levels of contaminating and extracellular microbes should be similar across cell-types because these cells are processed identically. In contrast, microbes that are enriched in one cell-type compared to the others likely represent intracellular microbes that reside within that cell-type. We demonstrate that our approach works using both human-designed benchmark scRNA-seq datasets and cancer scRNA-seq datasets with previously reported intracellular microbes. Finally, we apply our approach to a large, recently published scRNA-seq dataset from lung cancer and identify four tumors where microbial taxa are enriched compared to immune cells. By comparing the transcriptomes of infected and uninfected tumor cells, we identify antimicrobial peptides such as *S100A9* and multiple immune response pathways including antigen processing and presentation to be significantly downregulated in these infected tumor cells compared to the uninfected tumor cells.

Chapter 2: Modeling mutation signatures using clinical and molecular covariates

2.1 Preface

In this project, we develop an approach to flexibly incorporate tumor metadata into the mutation signature model and demonstrate that our approach can better distinguish between mutations caused by similar signatures. We focus on the problem of distinguishing between similar mutation signatures for two reasons. First, there is a well-established connection between topic modeling and mutation signature extraction from which we are able to borrow techniques. Second, there are important implications for precision medicine: mutation signature 3 has been proposed as a biomarker for PARP inhibitors but it can be difficult to distinguish mutation signature 3 from the similar mutation signature 5 using existing approaches.

A manuscript describing this project was accepted to the 2019 conference on Intelligent Systems for Molecular Biology (ISMB) in Basal, Switzerland where I presented this work. The manuscript was subsequently published in the journal *Bioinformatics* as part of the conference proceedings[41]. The approach and the design of the experiments in this project were conceived jointly by my co-advisor, Max Leiserson, and me. I wrote the code to perform all the experiments and analyze the results with help processing the TCGA datasets from Mark Keller. With the exception of Figure 2, which was generated with help from Mark Keller and Jason Fan, I generated the figures myself. Max Leiserson supervised this project.

2.2 Introduction

Somatic mutations accumulate over time in normal and cancer cells as a consequence of multiple mutational processes. Measuring and understanding the activity of these mutational process within and across tumors has important applications in modeling tumorigenesis, personalized cancer therapy, early detection, and prevention. The large cancer sequencing datasets generated over the past decade have led to the discovery of signatures of mutational processes present in patterns of single base substitutions[28]. Discovering and characterizing these *mutation signatures* and their underlying etiology has thus become an important challenge in the field.

The sources of somatic mutations can be broadly classified as due to errors in DNA replication or from environmental or lifestyle exposures[59]. Errors in DNA replication result both from processes active in healthy cells (e.g., due to spontaneous deamination or reactive oxygen species) and from perturbed DNA damage repair pathways[60]. Clinicians use measures of DNA damage repair deficiency for multiple types of cancer therapy, including chemotherapy, synthetic lethal therapy, and, more recently, checkpoint inhibitor immunotherapy[30,61,62]. A recent study evaluated mutation signatures of homologous recombination repair deficiency in breast cancer as a predictive biomarker, and found that the mutation signature-based approach would significantly expand the population of patients eligible for PARP inhibitors[63]. Mutations also result from environmental or lifestyle exposures, including UV radiation and tobacco smoke, as well as many DNA damaging agents used as chemotherapies[64,65].

Mutation signatures of these exogenous processes have recently been shown to be prognostic in cutaneous melanomas, and revealed pre-cancerous neoplasms resulting from aflatoxin B1 exposure[66,67]. More generally, these sources of somatic mutations can be thought of as tumor-level covariates where for a given covariate (e.g., smoking status), each tumor is annotated with a specific value (e.g., smoker or non-smoker).

The most widely used methods for discovering mutation signatures are based on non-negative matrix factorization (NMF) of a mutation count matrix[33]. To identify signatures in a cohort of N tumors, single base substitutions are first grouped into 96 categories (based on the substitution and its surrounding 5' and 3' contexts), yielding an N-by-96 matrix M of mutation counts. Then, NMF is applied to decompose M into a N-by-K *exposures* matrix E and a K-by-96 *signatures* matrix P, and E and P are rescaled so that the rows of P sum to one. Each entry $E_{i,j}$ is interpreted as the number of mutations in tumor i generated by signature j, and $P_{k,j}$ is the probability signature k generates a mutation of category j. Alexandrov *et al.* applied this model to >7000 tumors from 30 different cancer types to identify 20 mutation signatures[28]. Alexandrov and colleagues have since expanded the set to include 30 validated signatures that are widely studied and available from the Catalogue of Somatic Mutations in Cancer (COSMIC)[68].

Since Alexandrov *et al.* first applied NMF to identify mutation signatures, researchers have developed additional NMF algorithms, and addressed the problem of inferring exposures in a cohort given a set of active signatures. Kasar *et al.* introduced the SignatureAnalyzer method that uses a probabilistic formulation of NMF and

automatically learns its rank K[69]. Fischer *et al.* and Rosales *et al.* both introduced algorithms for NMF that assume that the mutation counts are drawn from a Poisson distribution parameterized by multiplying factors with a Gamma prior[70,71]. Rosenthal *et al.* introduced several heuristics for computing the exposure matrix E given a signature matrix P, and Huang *et al.* extended this work to solve the problem optimally[72,73].

A handful of researchers have also considered a second type of approaches to inferring mutation signatures that leverages lessons from the natural language processing problem of *topic modeling*. Given a corpus of observed documents, each drawn from the same vocabulary, the goal of topic modeling is to infer latent topics (distributions over words) and to assign each word in each document to its underlying topic[74]. Most topic modeling approaches such as the standard latent Dirichlet allocation (LDA) introduced by Blei *et al.* are Bayesian and make the ``bag-of-words'' assumption that each word in a document is independent given its underlying topic[35]. Applying topic modeling to mutation signatures means interpreting tumors as documents, signatures as topics, and mutation categories as the vocabulary. Shiraishi *et al.* introduced the pmsignatures method that generalizes LDA to enable mutation categorizations that include more than one flanking base[34]. Funnell *et al.* used a multi-modal topic modeling approach to simultaneously analyze patterns in single base substitutions and structural variations in breast and ovarian cancers[39].

Despite this methodological progress, about half of the 30 validated COSMIC signatures have no known etiology. The current approach to mapping signatures to their underlying causes is to show statistically significant associations between

signature exposure and a clinical/demographic features (e.g., a history of smoking and COSMIC Signature 4) or molecular features (e.g., *BRCA1* inactivations and COSMIC Signature 3)[75,76]. Further, even for two signatures with known etiologies, it can be challenging to distinguish their respective exposures with existing methods if the signatures are similar. For example, COSMIC Signature 3 and Signature 5 are highly similar (cosine similarity of 0.83), but Signature 3 is associated with homologous recombination repair deficiency and Signature 5 is associated with age at diagnosis and genetic mutations in the nucleotide excision repair pathway[40,76–78].

We hypothesize that to overcome these challenges, methods for modeling mutation signatures and tumor-level clinical or molecular covariates are needed. To begin to address this challenge, we present the Tumor Covariate Signature Model (TCSM) to learn how observed tumor-level covariates change signature exposure. We show on simulated and real mutation datasets that, by modeling tumor-level covariates, TCSM outperforms existing NMF- and topic modeling-based approaches that are limited to using only a tumor's mutations as input. We find that the largest differences in performance come when inferring exposures of held-out tumors not used to infer signatures, and that these differences lead to improved performance in downstream analyses, including predicting DNA damage repair deficiency. TCSM is the first method to model mutation signatures and their tumor-level covariates in order to automatically infer signature etiology.

2.3 Methods

2.3.1 Tumor Covariate Signature Model (TCSM)

We present a probabilistic model of mutation signatures and their covariates that builds off of the well-studied area of topic modeling and the previously observed connection between topic modeling and mutation signatures[34,35,39,74]. Topic models are generative models for text data, and usually encode the "bag-of-words" assumption that words are independent given their underlying topics. The observed data for topic models are N documents w, where each document w_i consists of n_i words from vocabulary V such that $w_{ij} \in V, 1 \leq j \leq n_i$. Topic modeling seeks to uncover (1) K global latent variables β_k called *topics*, where each topic is a probability distribution over the vocabulary; and, (2) local latent variables including the K *topic mixing proportions* θ_i per document, and the assignment $z_{ij} \in \{1,...,K\}$ of each observed word w_{ij} to a topic. The most common topic modeling approaches such as latent Dirichlet allocation (LDA) by Blei *et al.*[35] are Bayesian, where both β_k and θ_i are multinomial distributions with Dirichlet priors.

In order to model mutational processes in cancer, we interpret tumors as documents, mutation categories as the vocabulary, signatures as topics, and signature *exposures* as topic mixings. Following earlier work, we categorize mutations into L = 96 mutation categories based on its base substitution (C:G>A:T, C:G>G:C, C:G>T:A, T:A>A:T, T:A>C:G, T:A>G:C) and the 5' and 3' flanking bases (four choices each) in the reference genome.

We present TCSM to allow observed tumor covariates to change the per tumor distribution θ_i of signature exposures (**Figure 2**). While there is a rich history of topic

modeling using document-level covariates, to our knowledge, this is the first time this work has been connected to mutation signatures[37,38,79]. Importantly, we do not model the generative process of the observed covariates, but instead take a conditional approach where the D observed covariates $\sim x_i$ of the i^{th} tumor change the prior distribution over the signature exposures θ_i. For example, an observed covariate could be a binary indicator for biallelic inactivation of a DNA damage repair gene. The model is flexible enough that the covariates can be any real valued number. The first element of $\sim x$ is always set to 1 to model the mean exposure of each signature.

More specifically, we follow the "topic prevalence" approach of the Structural Topic Model from Roberts *et al.* that combines Dirichlet-multinomial regression and the correlated topic model, and describe the model as it relates to mutation signatures[36,38,79,80]. The correlated topic model places a logistic normal prior on θ such that signature exposures can co-vary (correlate) and was previously used to analyze mutation signatures in breast cancer by Funnel *et al.*[39] The mean of the logistic normal is set for tumor i as $x\sim_i\Gamma$, where Γ is a $D\times(K-1)$ matrix of exposure-covariate coefficients. The full generative process for the TCSM for tumor sample i with n_i mutations is as follows:

$$\theta_i \sim \text{LogisticNormal}(\sim x_i\Gamma,\Sigma),\ (1)\ z_{ij} \sim \text{Multinomial}(\theta_i),\ 1 \le j \le n_i,\ (2)\ w_{ij} \sim \text{Multinomial}(\beta_{zij}),\ 1 \le j \le n_i.\ (3)$$

We place a hyperprior on the exposure-covariate coefficients $\Gamma = [\gamma_1;...;\gamma_{K-1}]$ where

$$\gamma_{d,k} \sim \text{Normal}(0,\sigma_k^2), \quad 1 \le d \le D, 1 \le k \le K-1,$$

and a Half-Cauchy(1,1) prior is placed on σ_k to weakly enforce regularization.

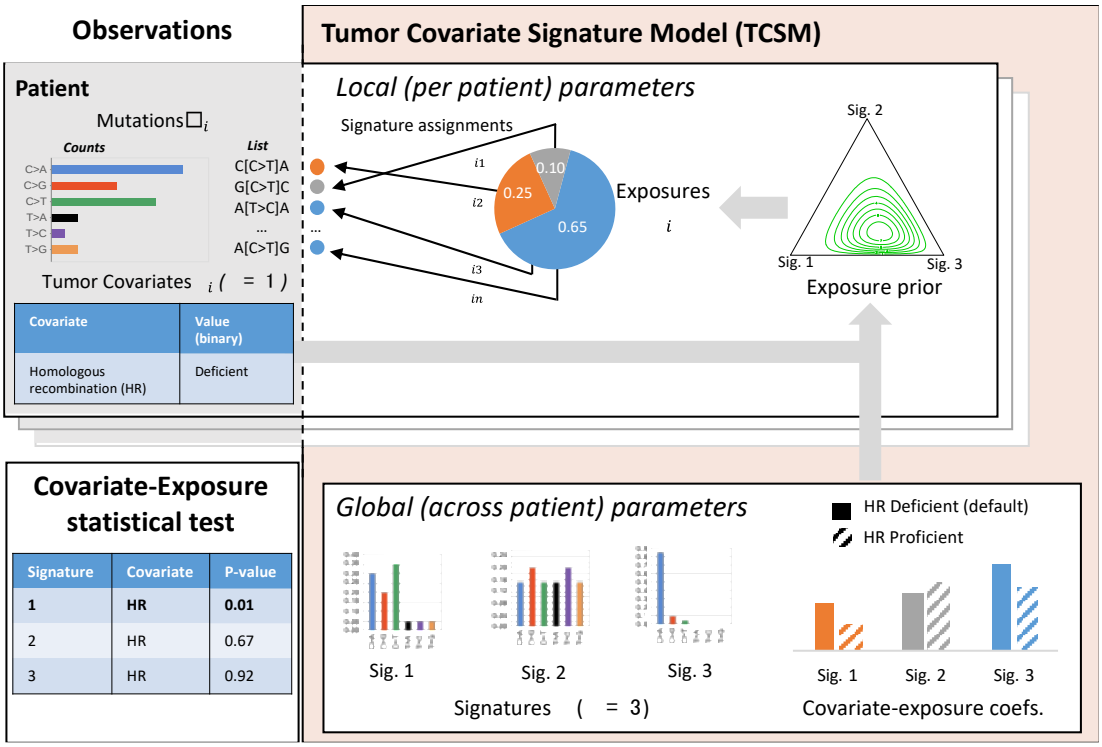


Figure 2: Overview of the Tumor Covariate Signature Model (TCSM)
We show an illustrative example of d=1 covariate and K=3 signatures. Given the observed mutations in a cohort of patients (top left), TCSM learns per patient exposures and assignments (top right), and a global set of signatures and covariate-exposure coefficients (bottom right). The associations between covariates and exposures are then tested for statistical significance (bottom left). Parts of the design of the figure are inspired by Blei et al. and Alexandrov et al.[33,74]

2.3.2 Model training and hyperparameter selection

We train the TCSM to learn the signatures β, signature exposures θ and covariance Σ, and exposure-covariate coefficients Γ using the variational expectation-maximization algorithm from Roberts *et al.* and their recommended initialization procedure[80]. The latter is based on a spectral decomposition (via non-negative matrix factorization) of the $L \times L$ mutation co-occurrence matrix that was shown to lead to quicker convergence of topic models[80].

The main hyperparameter of TCSM is the number K of signatures. We set K empirically through 5-fold cross-validation, completely holding out 20% of the

tumors in one fold. We use the "document completion" approach of Wallach *et al.* to compute the likelihood of all of a held-out tumors's mutations w_{test} i.e., computing $\Pr(w_{\text{test}}|\beta,\cdot)$, where $\cdot$ represents hyperparameters[81]. We choose the K when the likelihood plateaus.

Learning exposures in held-out samples. When the signatures β are given (e.g. from learning on a training cohort), we learn the exposures θ for additional, held-out samples by maximum a-posteriori probability (MAP) estimation.

2.3.3 Imputing binary covariates in held-out samples

One advantage of TCSM is that it enables probabilistic imputation of held out (or missing) covariates, including for previously unseen tumors. For example, for a single binary covariate in tumor x_{id}, we compute the loglikelihood ratio (LLR) of the tumor's mutations under the model with $x_{id} = 1$ and $x_{id} = 0$:

$$\text{LLR} = \log \frac{\Pr(w|x_{i1=1},\beta,\Sigma,\Gamma,\cdot)}{\Pr(w|x_{i1=0},\beta,\Sigma,\Gamma,\cdot)}, \qquad (5)$$

where $\cdot$ is the hyperparameters of the model. A positive *LLR* indicates that the tumor's mutations are better fit when $x_{id} = 1$. After imputing held-out or missing covariates in this way, we then report the exposures θ estimated from the model with higher likelihood for downstream analysis.

2.3.4 Statistical Significance of covariates on signature exposure

After applying variational EM to infer the latent variables of TCSM, we perform a statistical test for the significance of a covariate with respect to signature

exposure. In this work, we only perform the test for a single binary covariate. For each signature *k* and binary covariate *d*, we generate 10,000 random exposures to signature *k*, half setting $x_d = 1$ and half setting $x_d = 0$, according to Equation 1. We then generate an empirical distribution by repeating these steps for TCSM trained on data where the covariates are permuted among samples uniformly at random. We compute a *P*-value for a signature-covariate pair by counting how often the mean differences in exposure of any signature-covariate pairs on the permuted datasets are greater than the mean difference of exposures on real data. We specifically test for an increase in exposure and only report the cases where the mean exposure when the covariate is present is greater than the mean exposure without the covariate; the parameterization of the Dirichlet (or Logistic Normal approximation) necessarily means that increasing the exposure of one signature will decrease the exposure of at least one of the others. We report Benjamini Hochberg-corrected *P*-values[82].

2.3.5 Benchmarking of mutation signature methods

It is challenging to compare mutation signature methods on real data because the true signatures and exposures are unknown. For that reason, we perform comparisons on both simulated and real data.

2.3.5.1 Simulated mutation datasets

We generate simulated mutation datasets from a simplified version of TCSM with known ground truth parameters and hyperparameters based on real cancer datasets and previous mutation signatures studies. The simulation process is simplified in that we do not allow correlations between signature exposures, so

instead sample each tumor's exposures θ from a Dirichlet (as in Dirichlet-multinomial regression) instead of the logistic normal[38]. As a case study, we generate data to reflect homologous recombination (HR) repair deficiency in breast cancer, using a single binary covariate. We use four of the validated COSMIC signatures found to be active in breast cancer (Signatures 1, 2, 3 and 5)[68]. For each sample, we generate a single binary covariate x_i, representing HR deficiency, that increases the prior probability of exposure to Signature 3 (the COSMIC HR deficiency signature). We then generate θ_i of tumor i from a Dirichlet distribution with parameter vector $\eta_{ik} = \exp\{\lambda_{0,k} + \lambda_{1,k}x_i\}$. We use $\lambda_0 = [-2,-2,-5,-2]$ and $\lambda_1 = [0,0,4,0]$. Thus, simulated tumors with HR deficiency have a much greater prior probability of high Signature 3 exposure, while the other signatures prior probabilities remain unchanged. We note that Signatures 3 and 5 have a high cosine similarity of 0.83 to each other, making it challenging to distinguish between Signature 3 mutations resulting from HR deficiency and Signature 5 mutations.

2.3.5.2 Evaluation methods

To quantify the importance of tumor covariates in modeling mutation signatures, we compare the TCSM with and without covariates. We also compare the models to non-negative matrix factorization, using the popular Somatic Signatures implementation of NMF for mutation signature analysis[83].

Recovery of ground truth parameters. On simulated data, we compare the models on their learned signatures (using average cosine similarity) and exposures (using mean squared error). Note that these are in-sample comparisons.

Held-out log-likelihood. We compare TCSM with and without covariates using average log-likelihood per mutation of held-out data. Since NMF is non-probabilistic, we cannot compare it to TCSM using likelihood.

Prediction tasks using estimated exposures. To compare between probabilistic and non-probabilistic models, we compare the prediction power of the inferred exposures for a target binary covariate that is known to be associated with mutation signatures. First, we learn the mutation signature model on the training data set. Then, we use the model to estimate the exposures of the test data set to the identified signatures. Importantly, while the covariate is used when training TCSM, we hold it out completely in the testing dataset. For TCSM, we first impute the covariate in held-out samples before computing exposures (as described in Section 2.3.2). For NMF, we estimate the exposures in held-out samples using SignatureEstimation[72]. Next, a Support Vector Classification (SVC) model with a linear kernel is trained using the normalized exposures of the training dataset and the target covariate and evaluated on the test dataset. When the distribution of the target covariate is unbalanced, we set the class weight parameter of the SVC method to balanced and evaluate the performance using area under the precision-recall curve (AUPRC).

2.3.6 Implementation and software

We implemented TCSM in Python 3.We perform model training and inference using a wrapper of the Structural Topic Models R package[79]. We provide a workflow for reproducing the experiments in the paper using Snakemake[84]. The source code is publicly available at https://github.com/lrgr/tcsm.

2.3.7 Data

We analyze mutations in breast cancer exomes processed and standardized by The Cancer Genome Atlas PanCanAtlas and downloaded from the Genomic Data Commons[85]. To investigate the relationship between breast cancer and homologous recombination (HR) repair deficiency, we restrict our analysis to 760 tumors with called biallelic inactivation of 82 genes in the HR pathway and counts of LST (Large-scale State Transitions; a measure of HR deficiency) from Riaz *et al.*[31,86]. We obtain biallelic inactivation calls for the 82 HR genes by combining epigenetic silencing calls from Knijnenburg *et al.* with germline and somatic mutation and loss of heterozygosity (LOH) calls from Riaz *et al.*[31,87].

We also analyze 466 melanoma exomes and 485 lung squamous cell carcinoma tumors from The Cancer Genome Atlas PanCanAtlas dataset[85]. We exclude 48 melanoma samples that were annotated as either acral melanomas or metastatic samples with unknown primary tumor origin by [66] (list of excluded samples obtained via personal correspondence). We download CC>TT dinucleotide polymorphism counts for these samples from both Firehose and Alexandrov *et al.*[88] We combine these data sources by taking the average CC→TT count for samples that appear in both sources.

2.4 Results

2.4.1 Comparison on simulated data

We first compare the Tumor Covariate Signatures Model (TCSM) on simulated data with known ground truth to two baseline methods: nonnegative matrix

factorization (NMF) and TCSM using no covariates. To better understand how a single signature with changes in exposure due to tumor covariates affects the performance of TCSM and existing methods, we perform this comparison using simple simulated datasets with a single binary covariate that changes the prior probability of exposure for a single signature. The remaining parameters are set using previously discovered mutation signatures or are derived from real mutation datasets.

We randomly generate fifty simulated datasets (see Section 2.3.5), varying the number of samples from 50 to 250 and sampling with replacement the number of mutations per sample from real breast cancer exomes from The Cancer Genome Atlas PanCanAtlas dataset[85]. We then compare the output of our model to NMF as implemented by the SomaticSignatures R package[83]. We apply TCSM with and without covariates to directly quantify the importance of incorporating tumor covariates. We evaluate the models in terms of the log-likelihood of held-out samples for $K = 2{-}8$. We compute the average held-out log-likelihood using Monte Carlo cross-validation with fifty train/test splits, holding out 20% of the samples. We also report each model's in-sample accuracy at identifying the hidden signature and exposure parameters.

In terms of model selection (identifying the true K), we find that TCSM with covariates consistently outperforms TCSM without covariates and SomaticSignatures. While none of the models are able to consistently learn the true number of signatures ($K = 4$) in datasets with only 50 samples, TCSM identifies the true K more often than the other methods (7/50 times compared to 2 and 1 for TCSM without covariates and SomaticSignatures, respectively). We used the residual sum-of-squares and explained

variances for model selection for SomaticSignatures, as suggested by the authors[83]. When we use 250 samples, we find that TCSM with covariates identifies the true number of signatures ($K = 4$) for 35 of the datasets (compared to 3 and 19 for TCSM without covariates and SomaticSignatures, respectively). We also find that covariates provide additional signal, as TCSM with covariates achieves higher held-out likelihood than the TCSM without covariates on the majority of the synthetic datasets when $K = 4$ for $N = 50$ (28/50) and nearly all datasets when $N = 250$ (49/50). All models identified the signatures with relatively high accuracy (cosine similarity >0.90; **Figure 3A**) for $N > 100$. However, TCSM with covariates was better able to distinguish between mutations caused by Signatures 3 and mutations caused by Signature 5, with higher accuracy in identifying the true exposures across all datasets (**Figure 3B**).

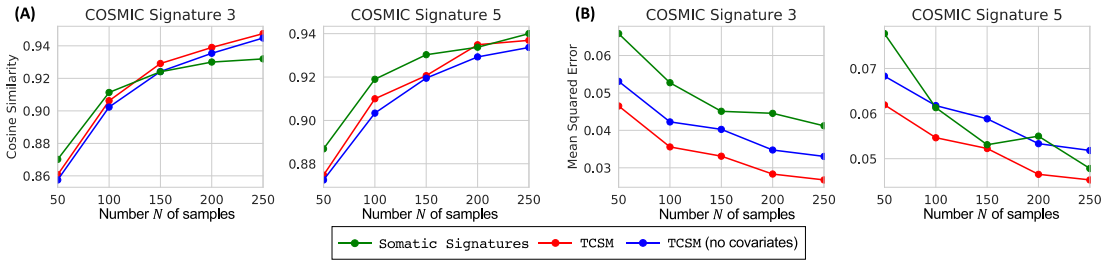


Figure 3: Benchmark of TCSM on simulated data
TCSM with (red) and without (blue) covariates is compared to the NMF-based SomaticSignatures (green) on synthetic data. ***(A)*** *Cosine similarity of inferred signatures (β) to hidden Signatures 3 and 5 using the true K = 4 averaged across fifty datasets, varying the number of samples.* ***(B)*** *Mean-squared error of the inferred exposures (θ) for the same datasets as in (A)*

2.4.2 Homologous recombination repair (HR) deficiency in breast cancer

After establishing the utility of our model on simulated data, we turn to test it on real data. As an initial case study, we apply TCSM to study homologous recombination (HR) repair deficiency in breast cancer. Understanding HR deficiency

in breast cancer is particularly important because of the clinical importance of identifying patients who might respond to PARP inhibitors[30]. We use the TCGA BRCA cohort and divide the samples stratified by the biallelic HR covariate (described below) into (1) a training/validation data set (75%) for choosing the encoding of the covariate, model selection, and benchmarking TCSM with/without covariates; and (2) a completely held-out test dataset (25%) for evaluation with a prediction task.

2.4.2.1 Covariate selection

The first key challenge in applying TCSM to real data is choosing the events or measures to use as covariates. Ideally, the covariates should be associated with changes in signature exposure and be easy to interpret biologically in order to reveal signature etiology. We begin by examining traditional markers of homologous recombination deficiency, including the biallelic inactivation of specific genes in the HR pathway[31] and the number of large-scale state transitions (LST), which are chromosomal breakages that generates fragments of at least 10 Mb[86].

We first compare TCSM using LST count to TCSM using the biallelic inactivation of a gene in the HR pathway as covariates in terms of held-out log-likelihood for $K = 2 - 10$ (Appendix A Figure 1). We encode the biallelic inactivation of a gene in the HR pathway as a single binary covariate where a 1 indicates the tumor has a biallelic inactivation in one of the seven genes (*ATM*, *BRCA1*, *BRCA2*, *CHEK2*, *FANCM*, *FANCF*, *RAD51C*) in the HR pathway inactivated in at least five samples in our cohort. We find that LST gives consistently better performance as measured in held-out log-likelihood, which makes intuitive sense as it is designed to

be a direct readout of the functional status of the HR pathway. However, even though TCSM can use continuous variables as covariates, binary covariates – such as whether a gene has a biallelic inactivation – are more interpretable and easier to analyze downstream, e.g., when inferring the true value in a previously unseen sample. Therefore, we search for a subset of the HR genes whose biallelic inactivation maximizes the mutual information with the number of LSTs. More specifically, we use a greedy algorithm that adds the HR gene whose inactivation maximizes the mutual information with LST, halting when the mutual information stops increasing. The genes in the identified set, *BRCA1*, *BRCA2* and *RAD51C*, exhibit almost perfect mutual exclusivity (1/57 tumors have co-occurring mutations), a pattern expected for genes in the same pathway[89]. Further, TCSM trained using a single covariate for these three genes achieves superior performance than TCSM trained using a single covariate for all seven genes and nearly the same performance as TCSM using LST count as the covariate (Appendix A Figure 1). In subsequent sections, we refer to TCSM with a single covariate – the biallelic inactivation of either *BRCA1*, *BRCA2* or *RAD51C* – as TCSM with the biallelic HR covariate.

2.4.2.2 Automated discovery of mutation signatures of etiology

After selecting the covariate to use, we perform model selection over the range $K = 2 - 10$ using the TCSM with the biallelic HR covariate. We select $K = 5$ as that is where the held-out log-likelihood plateaus and show the resulting signatures in Figure S3. All five signatures have cosine similarity > .8 to COSMIC signatures with known etiologies (Appendix A Figure 2); specifically, TCSM Signature 1 maps to the

APOBEC signatures (COSMIC Signatures 2 and 13), TCSM Signature 2 maps to the HR deficiency signature (COSMIC Signature 3), TCSM Signature 3 maps to the polymerase epsilon signature (COSMIC Signature 10), TCSM Signature 4 maps to the mismatch repair (MMR) deficiency signature (COSMIC Signature 6) and TCSM Signature 5 maps to the aging signature (COSMIC Signature 1). Reassuringly, our covariate significance test identifies statistically significant increases in exposure to one TCSM signature, the TCSM signature that resembles COSMIC Signature 3, in the presence of the biallelic HR covariate (HR-proficient mean: .200, HR-deficient mean: .418, Benjamini-Hochberg-corrected $P < .001$).

Next, we evaluate the ability of the TCSM to impute a hidden biallelic covariate value given a held-out tumor's mutations. We impute each tumor's biallelic covariate when it is in the test fold during 5-fold cross validation. The log-likelihood ratio of tumors with inactivation of known HR genes – including inactivation of one of the three HR genes used in training (orange) or four other HR genes (green) – is significantly greater than the ratio of the samples without the inactivation of known HR genes (blue; **Figure 4C**; Wilcoxon rank sum $P = 7e^{-22}$). Moreover, the tumors predicted to be HR deficient (i.e., those with LLR > 0) without known HR inactivation have a significantly higher number of LSTs than the tumors predicted to be HR proficient (Wilcoxon rank sum $P = 8e^{-10}$, **Figure 4C**), possibly indicating that they may have some form of HR deficiency due to some other event. Together, these results demonstrate the use of TCSM for automated discovery of mutation signatures and their etiology.

2.4.2.3 Comparison to other methods

We compare the performance of TCSM with the biallelic HR covariate to TCSM without covariates (Figure 4A) for $K = 2{-}10$. We find that using covariates leads to an increase in held-out log-likelihood for all $K > 2$.

Next, we add NMF to the comparison. Since NMF is not probabilistic, we compare the estimated exposures of the three methods. We use the SomaticSignature R package implementation of NMF using the SomaticSignatures model selection process. We choose $K = 5$ because the model selection yields a range from $K = 3 - 6$ (Appendix A Figure 4) and $K = 5$ enables the fairest comparison between the models. The five signatures extracted by SomaticSignatures map with cosine similarity > .8 to the same five COSMIC signatures as TCSM.

We compare how well the estimated exposures of each method for *held-out* tumors correspond with standard measures of HR deficiency. We train a linear model to classify tumor HR deficiency from the tumor's signature exposures. Davies *et al.* recently demonstrated the potential of a similar approach using NMF-based exposures to expand treatment with PARP inhibitors to a broader class of patients[63]. As ground truth HR deficiency, we use the biallelic inactivation of *BRCA1*, *BRCA2* or *RAD51C*. We then train the model on exposures from TCSM with the biallelic HR covariate, TCSM without covariates and SomaticSignatures (see Section 2.3.5.2 for details). To enable a fair comparison, TCSM is not provided with the true value for the biallelic HR covariate for the held-out tumors but instead infers the covariate value before estimating the exposure (see Section 2.3.3). We evaluate the models in terms of the

area under the precision-recall curve (AURPC) on held-out cohorts not used when training the classifier.

We first compare within the cross-validation framework used for model selection. TCSM with the biallelic HR covariate (mean AUPRC=.62 across the 5-folds) outperforms both TCSM without covariates (mean AUPRC=.57) and the NMF approach (mean AUPRC=.56). We then compare on the completely held out 25% samples not used for model selection or choosing the encoding for covariates. Again, we find that TCSM with the biallelic HR covariate (AUPRC=.64) outperforms both TCSM without covariates (AUPRC=.59) and the NMF approach (AUPRC=.58).

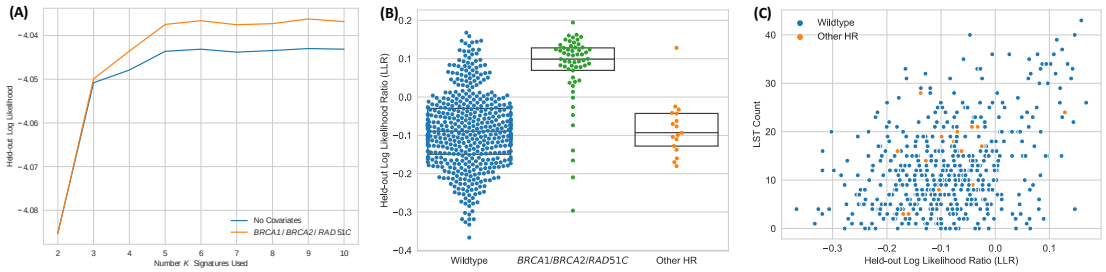


Figure 4: Evaluation of TCSM on TCGA breast cancer data
(A) *Comparison of the log-likelihood of held-out samples across K = 2 - 10 between TCSM with the biallelic HR covariate (inactivation of BRCA1, BRCA2, or RAD51C) and TCSM without covariates.* ***(B)*** *The log-likelihood ratio (LLR) of samples with the biallelic HR covariate hidden where LLR > 0 indicates the mutations of a sample are more likely under the biallelic HR covariate inactivation model.* ***(C)*** *After excluding tumors with known biallelic inactivation of either BRCA1, BRCA2 or RAD51C, the plot of a tumor's LLR against its LST count.*

2.4.3 Simultaneously learning signatures in melanomas and lung cancer

Next, we investigate mutation signatures in cutaneous melanomas (SKCM) and lung squamous cell carcinomas (LUSC), two types of cancer where mutational processes relating to environmental or lifestyle exposures are predominant. We examine whole-exome sequences of 418 SKCM and 485 LUSC tumors from TCGA

PanCanAtlas (see Section 2.3.7 for details). One advantage of TCSM is the ability to encode cancer type in the model while performing a pan-cancer analysis. In contrast, previous work searched for a consensus set of signatures from a pan-cancer run and individual cancer type runs[28,88].

We investigate using multiple covariates for TCSM: cancer type, smoking history (expected for many lung cancers), and exposure to UV radiation (expected for many melanomas). For cancer type, we use one binary covariate for SKCM and one binary covariate for LUSC. For smoking history, we set to one if the patient has a history of smoking and zero for never-smokers. Note that smoking history data are missing for SKCM patients, so we set their history of smoking covariates to zero. For UV radiation, we use the number of CC>TT mutations in the tumor, which has long been known as a marker of UV radiation exposure[90]. Note that these dinucleotide mutations are excluded from the traditional 96 single base substitution categories analyzed by mutation signature methods and are thus not included in the observations.

We first perform model selection using TCSM and compare the held-out log-likelihood using all four covariates (cancer type, smoking history and UV radiation exposure), using only the cancer type and using no covariates (**Figure 5A**). We find that using the cancer type covariates results in a large improvement in held-out likelihood across *K* compared to using no covariates (**Figure 5A**). In contrast, we find that using all four covariates results in a much smaller improvement in held-out likelihood across *K* compared to using only cancer type. We hypothesize that the additional covariates yield minimal improvement because they are strongly associated

with the cancer type. To simplify downstream analysis, we remove the smoking status and UV radiation exposure covariates and use only the cancer type covariate. To further simplify the model, we use a single cancer type covariate with two possible values (LUSC and SKCM), instead of using one binary covariate for each cancer type as these two models have identical held-out likelihood performance (Appendix A Figure 5). Using TCSM with the single cancer type covariate, we select $K = 4$ as the optimal number of signatures and show the resulting signatures in Appendix A Figure 7.

The four extracted signatures resemble known COSMIC signatures (Appendix A Figure 6): the ultraviolet (UV) radiation-associated signature (Signature 7), the smoking-associated signature (COSMIC Signature 4), the APOBEC-associated signature (Signatures 2 and 13) and a signature that resembles both the aging-associated signature (Signature 1) and the mismatch repair deficient signature (Signature 6), which is likely a composite of the two COSMIC signatures that share a high cosine similarity to each other (cosine similarity=.84). Reassuringly, TCSM finds an association between the SKCM cancer type and an increase in the exposure to the TCSM signature most similar to COSMIC Signature 7 (LUSC mean: .113, SKCM mean: .808, Benjamini Hochberg-corrected $P < .001$). TCSM finds an association between the LUSC cancer type and an increase in the smoking signature (LUSC mean: .448, SKCM mean: .054, Benjamini Hochberg-corrected $P < .001$), the APOBEC signature (LUSC mean: .180, SKCM mean: .013, Benjamini Hochberg-corrected $P < .001$) and the mismatch repair/aging signature (LUSC mean: .260, SKCM mean: .125, Benjamini Hochberg-corrected $P < .001$).

We then investigate imputing a tumor's cancer type from its mutations. Campbell *et al.* examined 660 lung adenocarcinomas (LUAD) and 484 LUSC from TCGA and identified three LUSC tumors whose molecular profile resembled melanomas[91]. They hypothesized that these three LUSC tumors might represent metastases from the skin and noted that one of these patients was previously diagnosed with basal cell carcinoma. Campbell *et al.* reported a similar result in a targeted sequencing dataset, such that 35% of hypermutated lung cancers had high COSMIC Signature 7 exposure[92]. Motivated by these reports, we use TCSM to reexamine the TCGA LUSC tumors to quantify the probability each primary tumor was correctly classified as LUSC.

We find that the cancer types imputed by TCSM are the same as the classified cancer type in the vast majority of cases (**Figure 5B**). All but three LUSC have negative log-likelihood ratios, and the three outliers all have LLRs > 1 (indicating that they strongly resemble melanomas). Indeed, these three outliers are the same as those Campbell *et al.* identified as having high UV radiation signature exposure. The number of CC>TT mutations in these tumors further supports the hypothesis that they are misclassified melanomas, as they are the only three tumors in the LUSC cohort with at least 15 CC>TT mutations (**Figure 5B**). This analysis confirms and expands upon the conclusions of Campbell *et al.* and demonstrates the use of TCSM for probabilistically reasoning about cancer type classification.

TCSM identifies several SKCM tumors as likely LUSC (LLR >0) that are less likely to be true misclassifications. One explanation is that SKCM tumors with LLR < 0 have very few mutations and almost no CC>TT mutations, especially when

compared to SKCM tumors with LLR > 0 (mean number of mutations: 70 vs. 1032, $P = 5e^{-27}$ Wilcoxon rank sum; mean number of CC>TT mutations: 0 vs. 23, $P = 1e^{-27}$). However, many SKCM tumors with very few or no CC>TT mutations are still correctly classified as SKCM tumors, which demonstrates the importance of using the entire mutation spectrum, instead of a single feature.

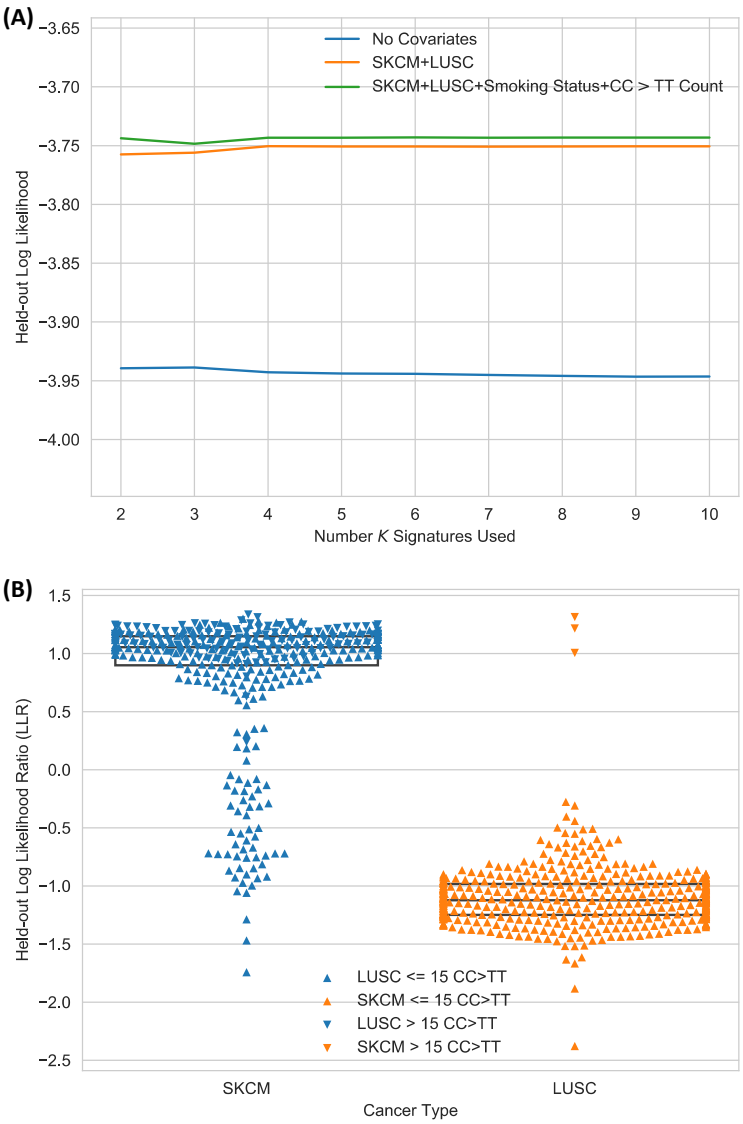


Figure 5: Evaluation of TCSM on TCGA melanoma and lung cancer samples
(A) *The held-out log-likelihood plot used for model selection to obtain K = 4.* ***(B)*** *The log-likelihood ratio (LLR) of the cancer type covariate for tumors where LLR < 0 means the*

mutations of the tumor are more likely under LUSC and LLR > 0 means the mutations of the tumor are more likely under SKCM.

2.5 Discussion

We presented the first probabilistic model, TCSM, of mutation signatures and their tumor-level clinical/demographic and molecular covariates. We found that TCSM outperformed NMF- and topic modeling-based approaches on both simulated and real mutation datasets, particularly in distinguishing between exposures of similar signatures. We then modeled mutation signatures of homologous recombination repair deficiency in breast cancers, demonstrating an approach for selecting interpretable covariates and predicting HR deficiency in held-out tumors. We also modeled mutation signatures in melanomas and lung cancers simultaneously. By including cancer type as a covariate, we were able to provide statistical support for earlier claims that three lung cancers in our cohort from The Cancer Genome Atlas are misdiagnosed metastatic melanomas.

The key advantage of TCSM over existing methods is in inferring exposures, particularly in distinguishing exposures of similar signatures. For example, we found that a linear model trained on exposures from TCSM was better able to predict HR deficiency than linear models trained on exposures from methods that do not model covariates. While not the focus of the applications in this study, we hypothesize that by modeling the effects of tumor covariates on signature exposures, TCSM may be more sensitive than existing methods in discovering rare signatures. To do so may require explicit modeling of the number of mutations per tumor.

While modeling tumor covariates of mutation signatures brings clear advantages, it also raises the challenge of encoding and selecting covariates for the

model. Encoding a particular covariate requires considering its sparseness and interpretability. Consider the covariate representing HR deficiency. We reasoned that the biallelic inactivation of genes in the HR pathway are more interpretable than existing HR indices – even if the HR indices may be a more direct encoding of the covariate – and that because the inactivation of each HR gene is sparse and approximately mutually exclusive, they could be combined into a single event. Selecting covariates also brings challenges, particularly when the mutational processes active in a cohort are not well understood, there are multiple covariates related to the same process, there is population structure or batch effects correlated with exposure, or for discovering new signatures. In this case, it may be important to add a covariate selection component to the model.

Certain aspects of TCSM are computationally expensive and can be improved. For example, choosing the value of *K*, the number of signatures, requires multiple runs of TCSM for each potential value of *K*. One future extension is to model *K* as a draw from a Dirichlet Process, a version of which is popular for topic modeling[93]. Another computationally expensive step is our statistical test, which requires sampling 10,000 random exposures from the model because the mean of the logistic normal distribution is parameterized by a vector of *K*-1 coefficients, which does not lend itself to an easy interpretation of the significance of exposure-covariate associations. Substituting the Dirichlet distribution for the logistic normal distribution, such as in Mimno and McCallum[38], would improve the direct interpretability of the parameters, which would enable a fully Bayesian approach for evaluating the significance of the exposure-covariate associations.

Finally, one direction we plan to explore in future work is modeling the effect of covariates on the signatures themselves, rather than their exposure. This is analogous to topic models of regional variation in language usage per topic[80,94,95]. There are multiple cases of researchers reporting multiple different signatures of the same mutational process, though it is not always clear what each of the distinct signatures represents. Learning how covariates change the signature themselves may help uncover these relationships.

Chapter 3: Identification of tumor-specific intracellular microbes from scRNA-seq using CSI-Microbes

3.1 Preface

In this project, we develop the first tool to identify cell-type specific intracellular bacteria from scRNA-seq data. We apply our tool to a large number of scRNA-seq datasets to identify tumor cell-specific microbes for two main reasons. First, it is computationally challenging to both identify microbial reads and then distinguish reads from contaminating microbes from those of true intracellular microbes, which we do by using different human cell-types as controls. Second, these tumor cell-specific intracellular microbes may be targeted by antibiotics or T cell-based therapy.

I built the pipeline and wrote the code to download the datasets, perform all the experiments and analyze the results. Alejandro Schäffer supervised experiments to improve the microbial read identification step of the pipeline and assisted in the administrative approval of the datasets. Fiorella Schischlik greatly assisted me in the generation of the figures. Eytan Ruppin supervised this project. We placed a version of this manuscript describing the initial version of the pipeline and the initial results from this project on the pre-print server bioRxiv in May, 2020[96]. This version of the manuscripts reflects a significant improvement and is currently being finalized for submission to a journal.

3.2 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[47–49,53,97,98]. *pks+ E. coli* have been shown to induce a mutation signature frequently found in colorectal carcinoma[99]. In pancreatic cancer, a subset of taxa from the class *Gammaproteobacteria* were shown to mediate tumor resistance to chemotherapy[100]. 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[58].

Recent papers have demonstrated that some members of the tumor microbiome live intracellularly in tumor and non-tumor cells within the tumor microenvironment. For example, the previously mentioned *Fusobacterium* has been shown to bind to ligands overexpressed by colorectal carcinoma cells; it can invade and exist intracellularly within these cells[47,101]. Another recent publication used multiple experimental techniques to interrogate the microbiome of seven cancer types and found that each cancer type has its own characteristic tumor microbiome and many intratumoral bacteria exist intracellularly in both tumor and immune cells[102].

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 can be exploited therapeutically[103]. Despite these advances, it is challenging 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. Just recently, the study by Nejman *et al.*[102], which characterized the composition of the tumor microbiomes using 16S ribosomal RNA[104,105] and identified the intracellular localization of some bacteria using staining, was unable 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**), aimed at identifying intracellular microbes that are *cell-type specific* from single cell RNA sequencing (scRNA-seq) datasets. Previous studies looking at microbial reads from scRNA-seq of host cells have generally focused on viruses[106,107]. The only previous study to analyze bacterial reads from scRNA-seq of host cells that we are aware of did so in the context of known *Salmonella* infection using a protocol designed to capture bacterial reads[108]. CSI-Microbes extends upon this approach by demonstrating that viruses and intracellular bacteria that preferentially reside within one cell-type can be identified from two commonly used scRNA-seq protocols (Smart-seq2 and 10x) without knowing *a priori* the infecting virus or bacteria as long as the microbe is represented in an input list of reference genomes. In this *de novo*

identification context it is necessary to consider all microbial reads identified from the datasets, many of which are likely contaminants. Using user-specified cell-type annotations (such as those based on host transcriptomic data), CSI-Microbes aims to identify microbial reads that are enriched in specific cell types. This step controls for contaminating and extracellular microbes, whose abundances is assumed not to vary significantly between cells of different types after proper normalization. Finally, we show that the microbial abundances of the intracellular microbes identified are likely to be of functional significance as they are associated with host transcriptomic changes.

We first test and validate our approach using two human-designed benchmark datasets where human immune cells were exposed to *Salmonella* and both infected and bystander cells underwent scRNA-seq[109,110]. To test CSI-Microbes in cancer, we analyze two 10x datasets from cancer types with previously reported tumor-specific intracellular microbes and show that it successfully identifies both the previously reported enrichment of *Merkel polyomavirus* in Merkel cell carcinoma cells and *Fusobacterium* in colorectal carcinoma cells as well as the novel enrichment of *Hathewaya histolytica* in colorectal carcinoma cells from one patient. Subsequently, we apply CSI-Microbes to analyze a Smart-seq2 dataset of ~11,000 cells from 13 lung tumors. We identify multiple bacterial taxa including the bacterial species *Cutibacterium acnes* in tumor cells of four lung tumors, the genus *Leptotrichia* in stromal cells of one lung tumor and multiple bacteria taxa in the immune cells of another lung tumor. Finally, we performed a differential expression analysis between tumor cells with and without sequence evidence for intracellular bacteria to identify

host transcriptomic changes associated with intracellular bacteria. Notably, we find the gene *S100A9,* which encodes half of the anti-microbial heterodimer calprotectin, to be the most down-regulated gene in tumor cells with intracellular bacteria. At the pathway level, we find that pathways associated with innate immune response (including defense response and response to biotic stimulus), antigen processing and presentation and multiple cytokine response pathways are downregulated in the infected tumor cells. These associations both testify to the significance of the results of CSI-Microbes and suggest potential mechanisms for how and why intracellular bacteria reside within tumor cells.

3.3 Results

3.3.1 Overview of CSI-Microbes

The inputs to CSI-Microbes are i) FASTQ files from scRNA-seq experiments and ii) cell metadata, including cell type annotations and iii) 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 abundance of the microbial taxa between cell types, and (b) *differential presence*, which compares the percentage of cells with at least one read from the microbial taxa between cell types. We use the differential presence test for sparsely populated 10x scRNA-seq datasets with few microbial reads and the differential abundance test, otherwise. The output is a list of candidate cell type-specific intracellular microbial taxa ranked by their differential abundance or presence.

The algorithm proceeds in the following steps (**Figure 6** and see Methods for a detailed description): **(1)** scRNA-seq reads are mapped to microbial genomes and spike-in transcripts (differential abundant test only) after filtering the host reads. **(2)** 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 log 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[111], are used to report the abundance of the microbial taxa to discriminate between cell types, for each microbial taxa. For the differential presence test, microbial read counts are compared across specific cell types using a two-sided binomial test and the statistical significance and effect size ($\log_2$ fold-change) are reported. 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[112]. **(3)** Post-hoc tests of contamination inspired by the decontam model[55] are performed using spike-in reads and empty wells if available (Methods). These include two tests, the spike-in 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 (Methods).

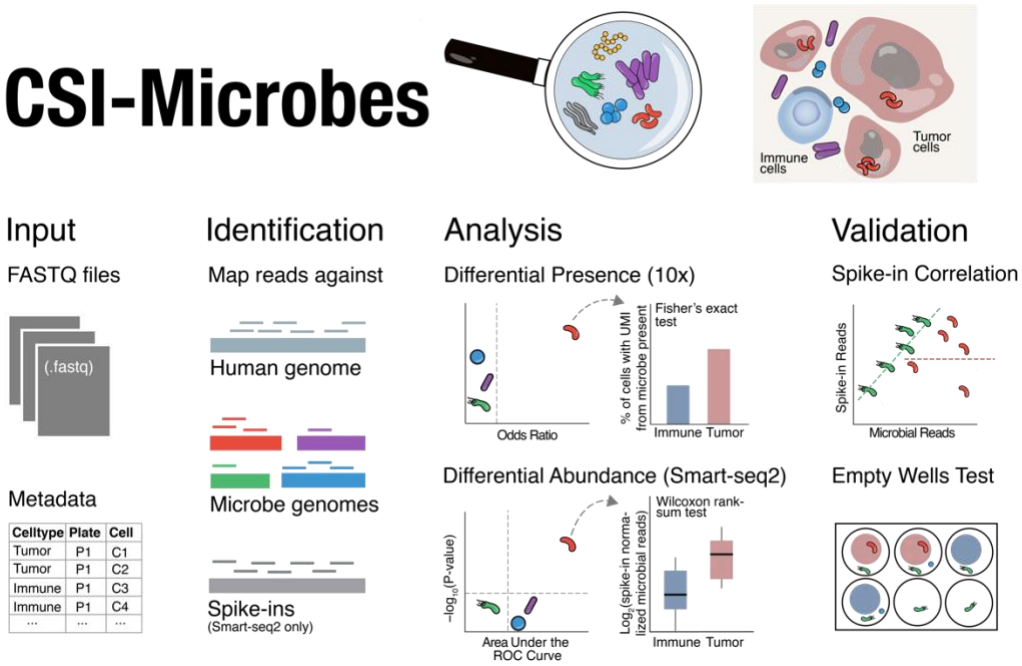


Figure 6: Overview of the CSI-Microbes approach
The expected input to CSI-Microbes is FASTQ files and metadata. The first step is (1) identification: the mapping of reads to human and microbial genomes and optionally spike-ins; (2) analysis: comparison of number of cells with at least one microbial UMI between cell-types (Differential Presence) or comparison of spike-in normalized microbial reads between cell-types (Differential Abundance); (3) validation: correlation of microbial reads with spike-in reads (Spike-in Test) and comparison between frequency of microbial reads in empty wells vs. wells with cells (Empty Wells Test)

3.3.2 Validation of CSI-Microbes on Salmonella exposed scRNA-seq datasets

We first test CSI-Microbes on a "gold-standard" Smart-seq2 dataset that sequenced 262 human monocyte-derived dendritic cells (moDCs) that had been exposed to either the *D23580* strain or the *LT2* strain of *Salmonella enterica* as well as 80 control "mock-infected" cells[109]. 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 using FACS. We identified a median of 8,030 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 the abundance of all but *Salmonella* to be highly correlated with the number of spike-in reads, suggesting that they are contaminants. We used CSI-Microbes to identify differentially abundant microbes between the infected and bystander cells and found only the taxonomic path from the class *Gammaproteobacteria* (p-value=$9e^{-6}$, AUC=.66) down to the species *Salmonella enterica* (p-value=$1e^{-8}$, AUC=.70) (**Figure 7A**). We did observe false positives when comparing cells across plates, illustrating the importance of controlling for the sequencing plate (step 2 in CSI-Microbes, Appendix B).

We next tested CSI-Microbes on a 10x dataset where the authors sequenced 3,485 human peripheral blood mononuclear cells (PBMCs) that were exposed to *Salmonella enterica* serovar Typhimurium strain *SL1344* as well as 3,515 unexposed control cells[110]. Using flow cytometry, the authors determined that ~3% of the exposed peripheral blood mononuclear cells (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* despite only 29 UMIs that mapped to bacterial genomes in this dataset (**Figure 7B**).

Although we do not find a significant difference between the percentage of genera-resolved bacterial reads belonging to *Salmonella* between the two datasets (8/27 vs. 756,284/1,643,696 Fisher Exact Test p-value=.12), we do observe significantly more microbial reads per cell in the Smart-seq2 dataset compared to the 10x dataset. Given this difference, we find it pertinent to employ different approaches

for filtering and false discovery correction for 10x vs Smart-seq2 datasets: For 10x datasets, we filter microbes that are not present above a minimum threshold of 1% in any cell-type, which filters all microbial taxa not belonging to *Salmonella* in the dataset from Bossel Ben-Moshe *et al*.[110] For Smart-seq2 datasets, we filter microbial taxa that have fewer than 10 counts per million microbial reads in at least 50% of the cells in any cell-type. We control for false discovery rate using hierarchical FDR, which leverages the ability of CSI-Microbes to identify differentially abundant taxa starting at the class taxonomic level in the NCBI Taxonomy and we report the highest resolution, statistically significant microbial taxa (Methods)[113,114]. We validated our findings of differential abundance and differential presence in the *Salmonella* datasets using direct mapping to the respective strain genomes using the error-tolerant aligner SRPRISM[115] (Appendix B).

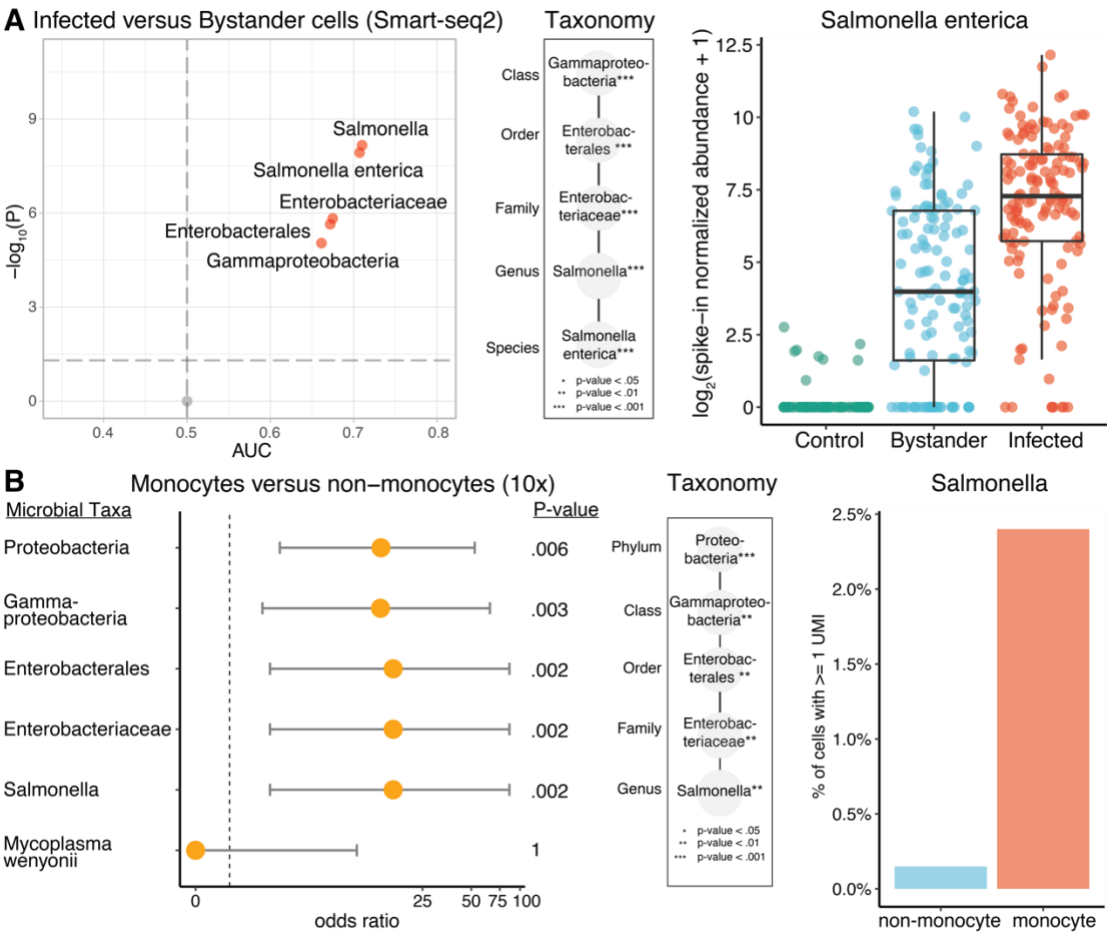


Figure 7: Results from CSI-Microbes on human cells exposed to Salmonella
(A) *Overview of the results from running CSI-Microbes on the dataset from Aulicino et al.*[109]*. The first plot is a volcano plot where all microbes were plotting according to the output of the differential abundance test (p-value and AUC). The second plot shows the taxonomic ordering of the differentially abundant microbes. The third panel shows the abundance of Salmonella enterica from infected, bystander and control cells.* ***(B)*** *Output of CSI-Microbes on the dataset from sample GSM3454529 from Bossel Ben-Moshe et al.*[110] *The first plot shows the output of the differential presence test (the taxonomic ancestors of Mycoplasma wenyonii received identical scores and were excluded for space purposes). The second plot shows the taxonomic ordering of the differentially present microbes. The third panel shows the frequency of the presence of Salmonella in monocytes and non-monocytes.*

3.3.3 Application of CSI-Microbes to Merkel cell and colon carcinomas

We further validate CSI-Microbes by analyzing two 10x scRNA-seq datasets from two tumor types that have previously been reported to have tumor cell-specific intracellular microbes. We first applied CSI-Microbes to identify different present

microbes between tumor and non-tumor cells from two Merkel cell tumors, which are among the ~80% of these cancers that are driven by the clonal integration of the Merkel polyomavirus[52,116]. 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: p-value=$6e^{-5}$, LFC=2.4; patient 9245-3: p-value=$3e^{-36}$, LFC=2.4).

Next, we applied CSI-Microbes to identify differentially present microbes between tumor and non-tumor cells from colorectal carcinomas, following previous reports that the bacterial species *Fusobacterium nucleatum* preferentially exists within colorectal carcinoma cells and to a lesser extent, stromal cells[47,117]. In agreement with these reports, CSI-Microbes identifies the genus *Fusobacterium* to be differentially present in the tumor cells from patient SC028 (**Figure 8A**). CSI-Microbes also identifies the differential presence of the bacterial species *Hathewaya histolytica* (previously called *Clostridium histolyticum*) in tumor cells of patient SC019 and with a trend towards enrichment in the tumor cells from patient SC030 (p-value=.24, LFC=2.33) (p-value=.009, **Figure 8B**). This species have been previously reported to be strongly enriched in the colonic tissue of patients with ulcerative colitis compared to controls[118].

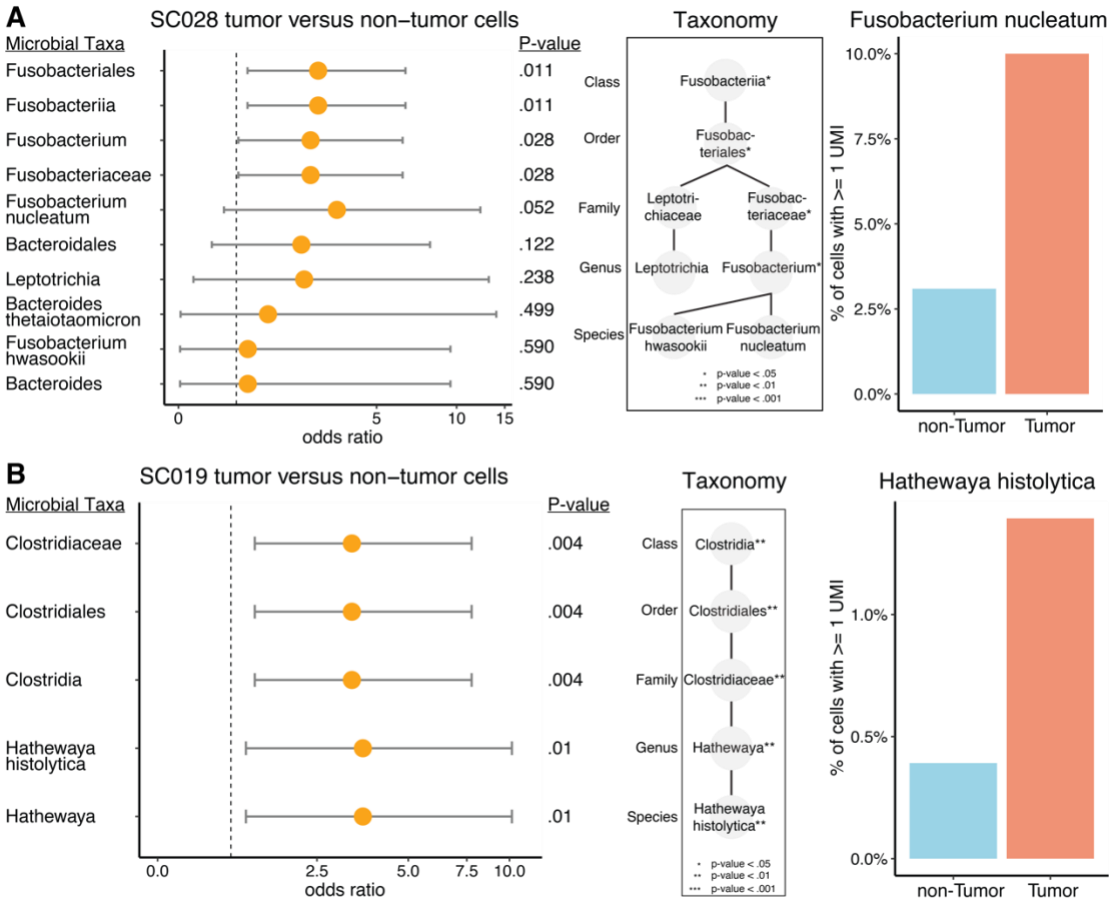


Figure 8: Results from CSI-Microbes on colorectal carcinoma
(A) *Output of the differential presence test of CSI-Microbes between the tumor and non-tumor cells from patient SC028 (non-significant ancestors are excluded for space reasons). The differentially present microbes and their children are ordered using the NCBI taxonomy. The percentage of cells with reads from Fusobacterium nucleatum are show between tumor and non-tumor cells* ***(B)*** *Output of the differential presence test of CSI-Microbes between the tumor and non-tumor cells from patient SC019. The differentially present microbes are ordered using the NCBI taxonomy. The percentage of cells with reads from Hathewaya histolytica are show between tumor and non-tumor cells*

3.3.4 Application of CSI-Microbes to lung cancer

Next, we applied CSI-Microbes to identify differential abundant microbes from a large, recently published lung cancer Smart-seq2 scRNA-seq dataset with spike-in sequences[119]. We analyze 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. Using the author's cell-type annotations

(tumor, immune, stroma and epithelial), we identify multiple tumors where microbial taxa are differentially abundant in tumor cells compared to immune cells (TH231, TH236, TH238, TH266, see examples in **Figure 9A and 9B**) and stromal cells (TH236, TH266). We also detect two tumors with taxa that are differentially abundant in stromal cells (TH231) or immune cells (TH220) compared to tumor cells. 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. Finally, comparing the results of CSI-Microbes to the results of 16S rRNA sequencing by Nejman *et al.*[102] in the lung, we find at least one unambiguous read to 16 of the 17 species found enriched in lung cancer by Nejman *et al.*[102], suggesting that scRNA-seq data may provide sufficient coverage of the tumor microbiome.

CSI-Microbes identifies the species *Cutibacterium acnes* to be differentially abundant in the tumor cells compared to the immune cells in four tumors (TH231, TH236, TH238, TH266). *Cutibacterium acnes* was excluded from a previous experimental exploration of the lung tumor microbiome by Nejman *et al.*[102] because it was identified in a large percentage of the negative controls, which indicated that it may be 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 also been reported as one of the most abundant commensals in the lung and to

exist intracellularly in epithelial cells[120,121]. Thus, unlike bulk expression based computational methods, CSI-Microbes can consider all microbes while implicitly controlling for contaminants by comparing between cells of the same patient. CSI-Microbes identifies another member of the *Cutibacterium* genus, *Cutibacterium granulosum* as differentially abundant in tumor cells in patient TH266 (uncorrected p-value = .04, **Figure 8B**). Additional genera that are differentially abundant in tumor cells include the genera *Corynebacterium* (TH236 and TH238) and *Staphylococcus* (TH236) and the family *Micrococcaceae* (TH238) (**Figure 8A**). 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 compared to other cells. In patient TH220, CSI-Microbes identifies both the *Micrococcus* and *Corynebacterium* genera to be enriched in immune cells compared to tumor cells in patient TH220. We do not find any bacterial taxa to be differentially abundant between macrophages/monocytes and any other cell type.

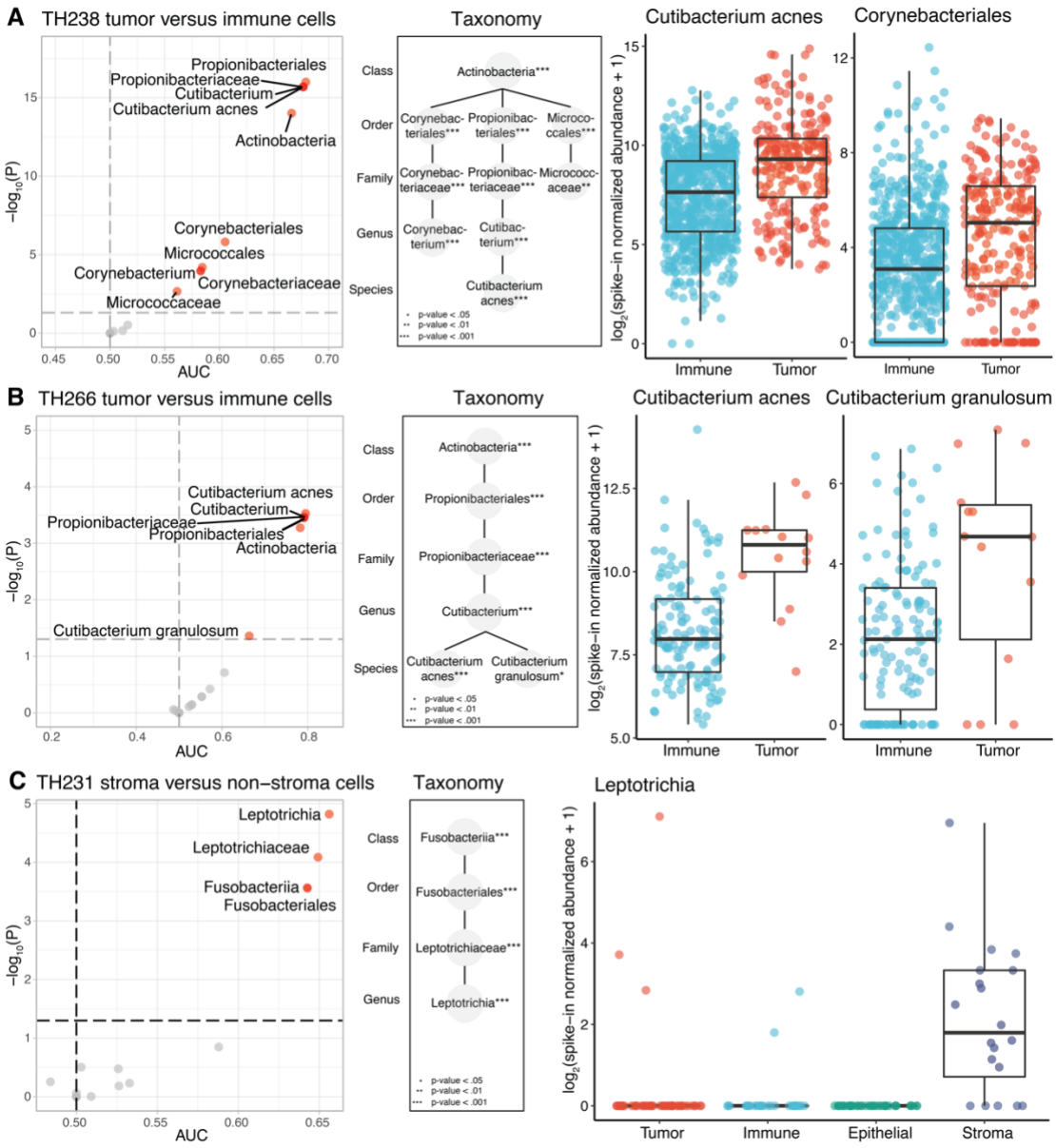


Figure 9: Results from CSI-Microbes on lung cancer

(A) *Overview of the differentially abundant microbes between the tumor and immune cells in patient TH238 including a volcano plot, the taxonomical relationship and the abundance of specific microbial taxa.* ***(B)*** *Overview of the differentially abundant microbes including both Cutibacterium acnes and Cutibacterium granulosum between the tumor and immune cells in patient TH266.* ***(C)*** *Overview of the differentially abundant microbes between the stroma and non-stroma cells in patient TH231 including the volcano plot and the taxonomic relationship as well as the abundance of genus Leptotrichia across the four major cell types in plate B003119 (the only plate from TH231 containing > 1 stroma cell)*

To study the transcriptomic state associated with the presence of intracellular bacteria, 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") (**Figure 10A**, Methods). At the gene level, the gene most down-regulated in infected tumor cells compared to uninfected tumors cells is *S100A9* (FDR-corrected p-value=$1e^{-62}$, AUC=.09), which forms a heterodimer calprotectin with *S100A8*. Calprotectin has antimicrobial properties because of its ability to sequester metal ions such as zinc, manganese and iron that are essential nutrients for microbes[122]. The strong down-regulation of calprotectin as well as multiple other S100 calcium-binding proteins may explain how bacteria such as *C. acnes* can survive inside 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 [123] (**Figure 10B**, Methods). The largest cluster of gene sets downregulated in infected tumor cells contains mostly gene sets associated with processing and presentation of antigens as well as gene sets 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. Of note, there are at least three additional and unconnected downregulated gene sets involved in anti-microbial response, including humoral immune response mediated by antimicrobial peptides, transition metal ion homeostasis and cell killing. Additionally, multiple immune response pathways such as response to interferon gamma and interferon beta

as well as interleukin-12 production are strongly downregulated in the infected tumor vs uninfected 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 [124]. The association of intracellular bacteria with the down regulation of the antigen presentation system in tumor cells, which both we (Appendix B) and Aulicino *et al.*[109] observe in the *Salmonella* dataset, 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[103].

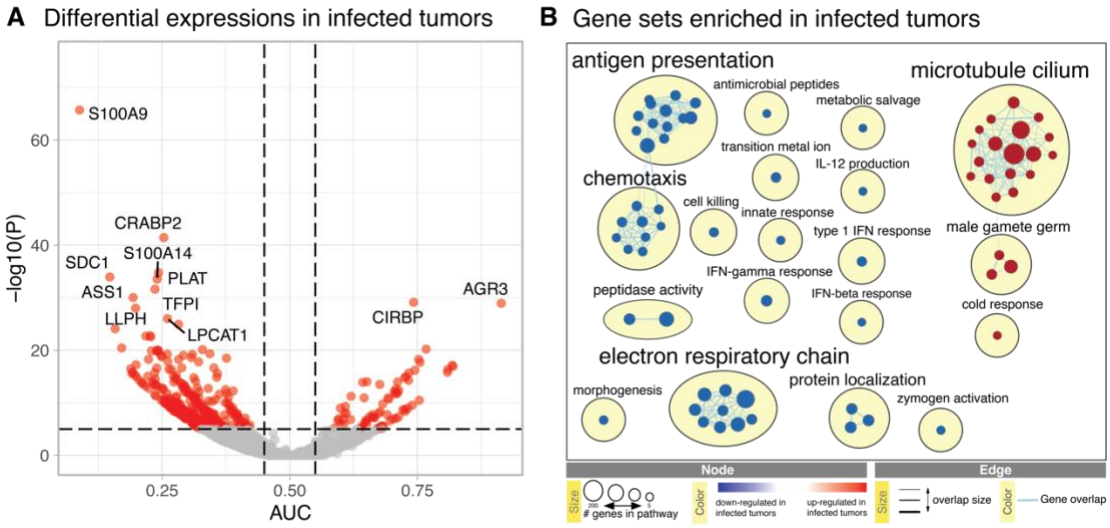


Figure 10: Transcriptomic changes between infected and uninfected tumor cells
(A) *Volcano plot of the differentially expressed genes between the infected and uninfected tumors cells.* ***(B)*** *Enrichment map of the enriched gene sets (FDR q-value < .02) where nodes represent gene sets and edges connect gene sets that share a high number of genes. Similar gene sets are clustered and manually named using common terms.*

3.4 Discussion

This paper introduces a new approach for the *de novo* identification of cell type-specific intracellular microbes from scRNA-seq data. We first demonstrate that CSI-Microbes can identify cells infected with intracellular bacteria from bystander

cells analyzing both 10x and Smart-seq2 scRNA-seq datasets and correctly identify the infective species. Next, we apply it to analyze scRNA-seq datasets from three different cancer types, showing that it identifies cell-type specific intracellular bacteria that have been previously reported in the literature, and additionally finds sequence evidence for the cell-type specific presence of other intracellular microbes, predominately in tumor cells but also in stromal and immune cells.

One limitation of this paper is that the commonly used scRNA-seq protocols that we analyze use polyA tail selection to enrich for polyadenylated eukaryotic mRNAs, which selects against prokaryotic RNA molecules, which are less likely to be polyadenylated and have shorter polyA tails[125]. Despite this under sampling, our computational approach finds clear evidence of the presence of bacterial reads in tumors in a cell-type specific manner. These findings call for further experimental testing and validation, e.g., using RNAscope[126] to learn how these genomic findings correlate with other means of detecting the presence of the microbes in cells. The recent findings of intracellular bacteria within tumor cells[102,103] obviously raises 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? Our findings that moDCs infected with intracellular *Salmonella* downregulate their antigen processing as already suggested by Aulicino *et al.*[109] point to a potential win-win relationship between intracellular bacteria and the tumors that host them, in which their presence leads to the down-regulation of the antigen processing and

presentation system of the host cell and supports tumor evasion of the immune system.

Finally, we note that CSI-Microbes can be applied to analyze any scRNA-seq dataset with multiple cell-types. 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.

3.5 Methods

3.5.1 Code and Data Availability

We analyzed publicly available FASTQ files from the following datasets:

scRNA-seq of monocyte-derived dendritic cells (MoDCs) exposed to *Salmonella* (Smart-seq2)[109]	BioProject PRJNA437328
scRNA-seq of PBMCs exposed to *Salmonella* (10x)[110]	BioProject PRJNA503437
scRNA-seq of Merkel cell carcinoma tumors (10x)[116]	BioProject PRJNA483959 (patient 2586-4), PRJNA484204 (patient 9245-3)
scRNA-seq of colorectal carcinoma tumors (10x)[117]	ArrayExpress E-MTAB-8410
scRNA-seq of lung cancer cells (Smart-seq2)[119]	BioProject PRJNA591860

The code is logically partitioned into two modules, one module for the "identification" step and one module for the "analysis" step. A reproducible Snakemake workflow for identifying microbial reads from scRNA-seq datasets, which includes the step of downloading the data from the datasets above, is available on GitHub (https://github.com/ruppinlab/CSI-Microbes-identification) although we

note that the identification module has some dependencies to the NIH Biowulf server. To facilitate reproduction of our analyses, we have uploaded the relevant microbial read abundance files to Zenodo. Using these files, the key results from this manuscript can be reproduced using a Snakemake workflow focused on the analysis module and available on GitHub (https://github.com/ruppinlab/CSI-Microbes-analysis).

3.5.2 Preprocessing Steps

3.5.2.1 Smart-seq2 datasets

Raw FASTQ files were trimmed using fastp v0.20.1 with the arguments "*--unqualified_percent_limit 40 --cut_tail --low_complexity_filter --trim_poly_x*"[127]. The trimmed FASTQ files were aligned to the reference human genome (GRCh38 gencode release 34) and any applicable spike-in sequences using STAR 2.7.6a_patch_2020-11-16 with the arguments "--soloType SmartSeq --soloUMIdedup Exact --soloStrand Unstranded --outSAMunmapped Within"[128].

3.5.2.2 10x datasets

Raw FASTQ files were aligned to the reference human genome using CellRanger v5.0.1 [17] (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.

3.5.3 Alignment of unmapped reads to microbial genomes

The unaligned reads were assigned to microbial genomes using PathSeq v4.1.8.1 with the arguments "*--filter-duplicates false --min-score-identity .7*"[54]. We constructed the reference microbial genome database by downloading the set of complete viral, bacterial and fungal genomes from RefSeq release 201[129]. 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 "gold-standard" experiments. 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[130].

3.5.4 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 (http://software.broadinstitute.org/pathseq/). The abundances are normalized using the computeSpikeFactors function from scran v1.16.0, which computes the library size factors using the sum of the spike-in sequences[131]. 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 normalized log abundances are compared across cell-types using the findMarkers function from scran v1.16.0 with arguments "*test='wilcox', lfc=0.5, block='plate'*". In scran v1.16.0, the two-sided p-

value ("*direction='any'*") when *lfc* > 0 is less unintuitively less conservative than the one-sided p-value (either "*direction='up'*" or "*direction='down'*") (https://github.com/MarioniLab/scran/issues/86) 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 suggested on p.79 by Sokal and Rohlf[132].

3.5.5 False Discovery Rate Correction

We use 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 OTUs tested at that taxonomic level using the Benjamini-Hochberg procedure [82]. For the CSI-Microbes results from the cancer datasets, we leverage 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[133] which implements the "outer-nodes" method of Yuketeli[113], which is the method in that theoretical paper that is 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. $2.54e^{-8}$).

3.5.6 Normalization Model

We extend the model used by decontam 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)[55]. 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[55]. 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.

3.5.7 Comparison to 16S Tumor Microbiome Findings

We compared our findings of presence of bacterial taxa as numerical identifiers in NCBI's Taxonomy tree[114] to the findings of Nejman *et al.*[102]. To do this comparison, we had to i) map the findings of Nejman *et al.*[102] 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 Nejman *et al.*[102], 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 species identified by Nejman *et al.*[102] 739 of these 1,783 species have at least one completely sequenced genome and were included in our microbial database. These species included 17 lung-cancer matches from Nejman *et al.*[102]

3.5.8 Gene Set Enrichment Analysis

To perform GSEA between the infected and uninfected tumor cells, we first performed differential expression analysis as describe 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 v4.1.0 with default settings and seed=149[134] with the gene ontology biological processes gene set from the molecular signature database (MSigDB) v7.3 [134–137]. We visualized the enriched gene sets (FDR q-value < .02) using Enrichment Map v3.3.1 (node cutoff FDR q-value < .02, edge cutoff similarity=.375).

Chapter 4: Conclusions

"If the 20th century was the century of physics, the 21st century will be the century of biology"

- Craig J. Venter and Daniel Cohen

If the 21st century is the century of biology (and early indications are positive), then there will have to be a significant amount of computational innovation. The ability to generate large amounts of more and more sophisticated biological data will only continue to grow but this data will only lead to biological discoveries if it can be properly analyzed. In this thesis, I present two very different computational approaches for the analysis of NGS data that illustrate two different types of computational innovations for the analysis of biological data.

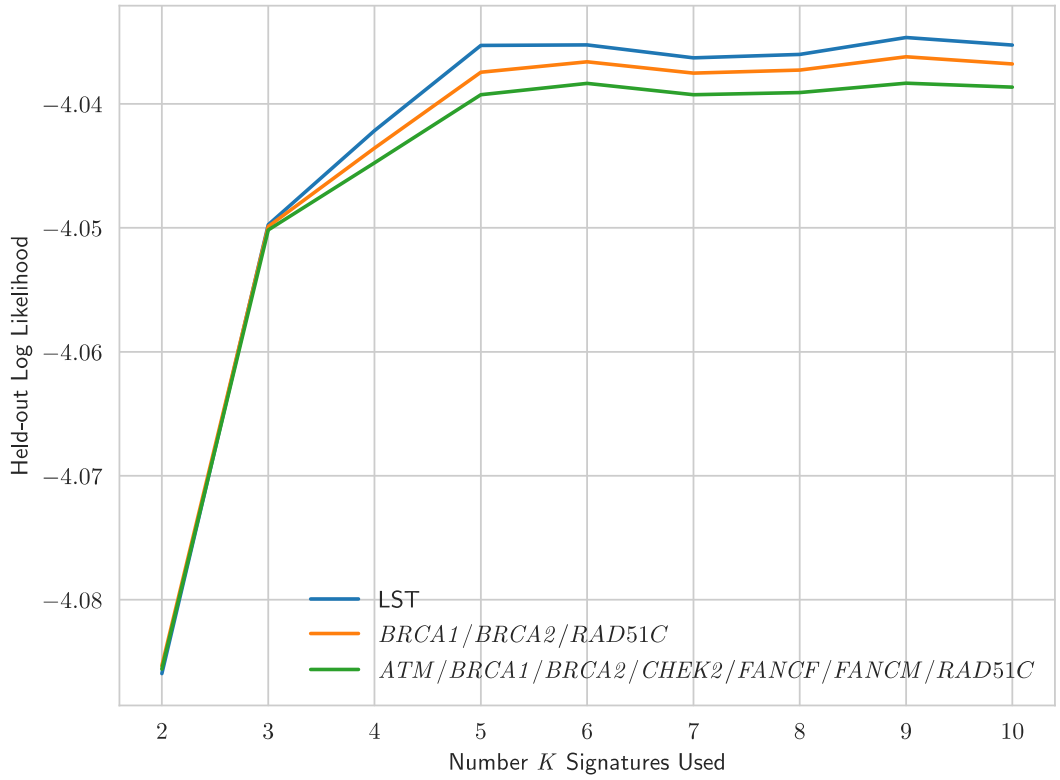
For TCSM, which was my first project, we applied a novel approach (borrowed from the field of topic modeling) to the well-studied problem of mutation signature extraction and were able to mainly rely on pre-processed data. For my second project, CSI-Microbes, we applied a novel approach to the under-studied question of the identification of intracellular microbes. To the best of our knowledge, CSI-Microbes is the first and only approach to use NGS to try to identify intracellular microbes. Our approach uses scRNA-seq data, which is a relatively new technology (in comparison to DNA sequencing). This project required both the software engineering ability to build scalable, efficient and reproducible pipelines to mine the raw reads of hundreds of thousands of cells (terabytes of data) for microbial reads but also the computational

ability to develop proper approaches for normalizing extremely sparse microbial reads between single cells.

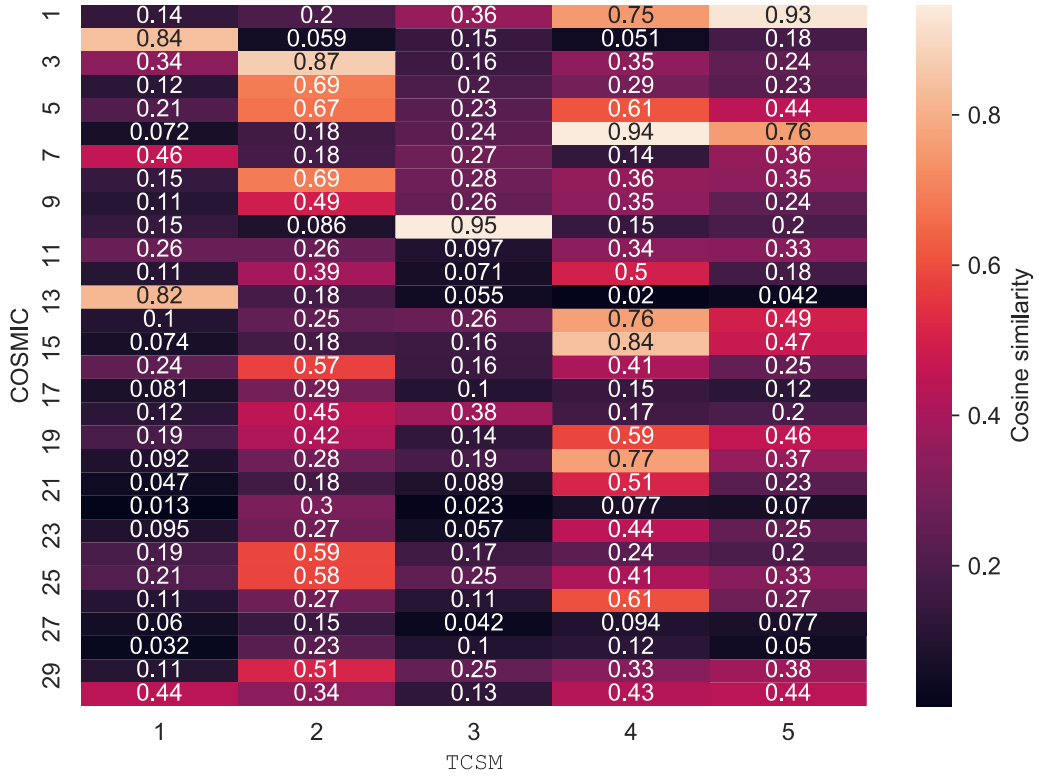
Importantly, both of these research projects were computational methods motivated by biological observations by myself and my co-authors. In particular, we were only able to identify the research question behind CSI-Microbes because of my collaboration with the Surgery Branch. One of my most important advantages as a computational biologist is that I am genuinely interested in learning and understanding the underlying biology. This curiosity has allowed me to both apply novel computational methods to solve research questions posed by others and pose my own novel research questions.

Appendices

Appendix A

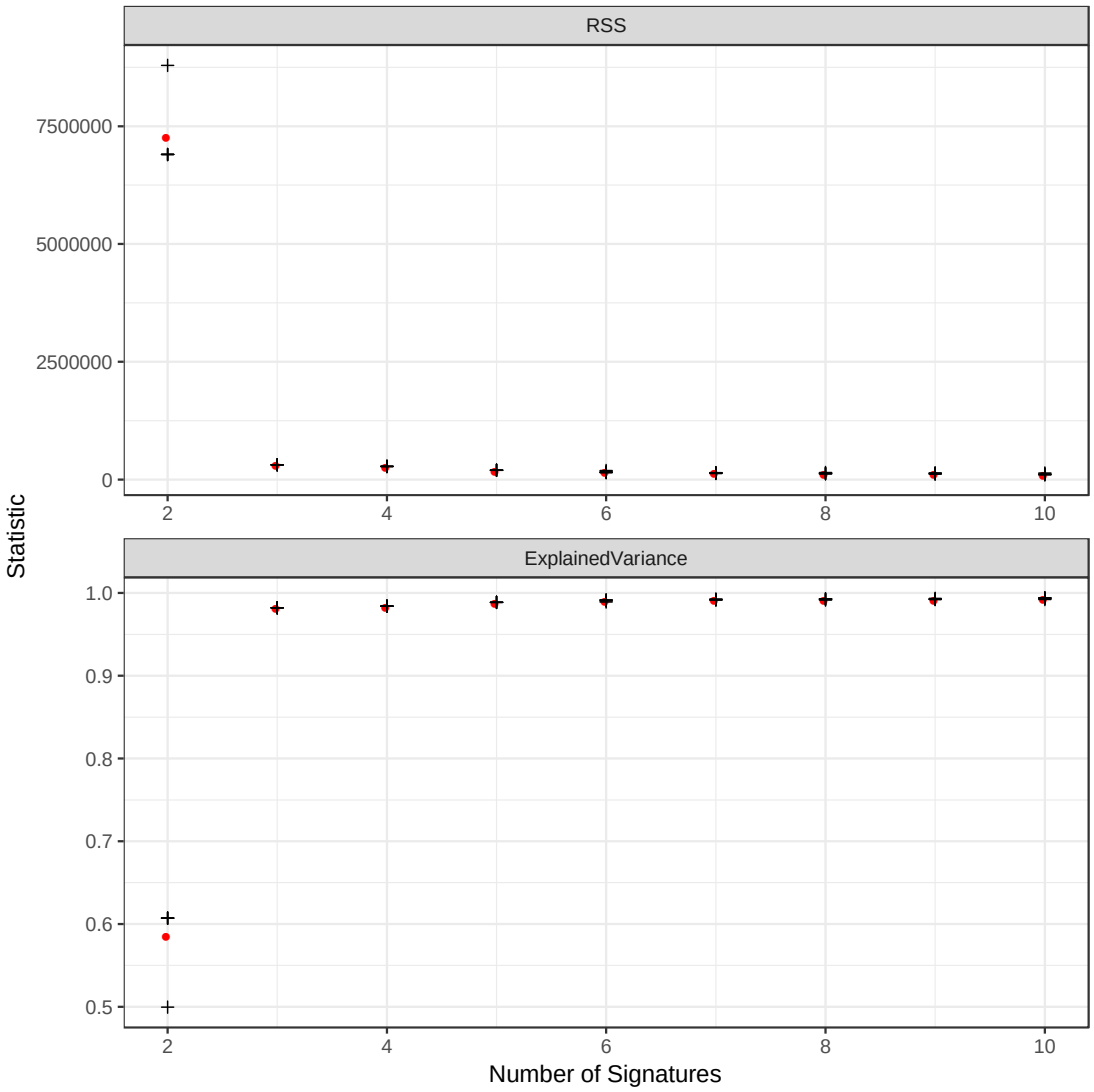


Appendix A Figure 1: Comparison of the likelihood of held-out samples of TCSM with the LST count, TCSM with the HR biallelic covariate (BRCA1, BRCA2, RAD51C) and TCSM with the biallelic inactivation of any of seven recurrently inactivated HR genes (ATM, BRCA1, BRCA2, CHEK2, FANCM, FANCF, RAD51C).

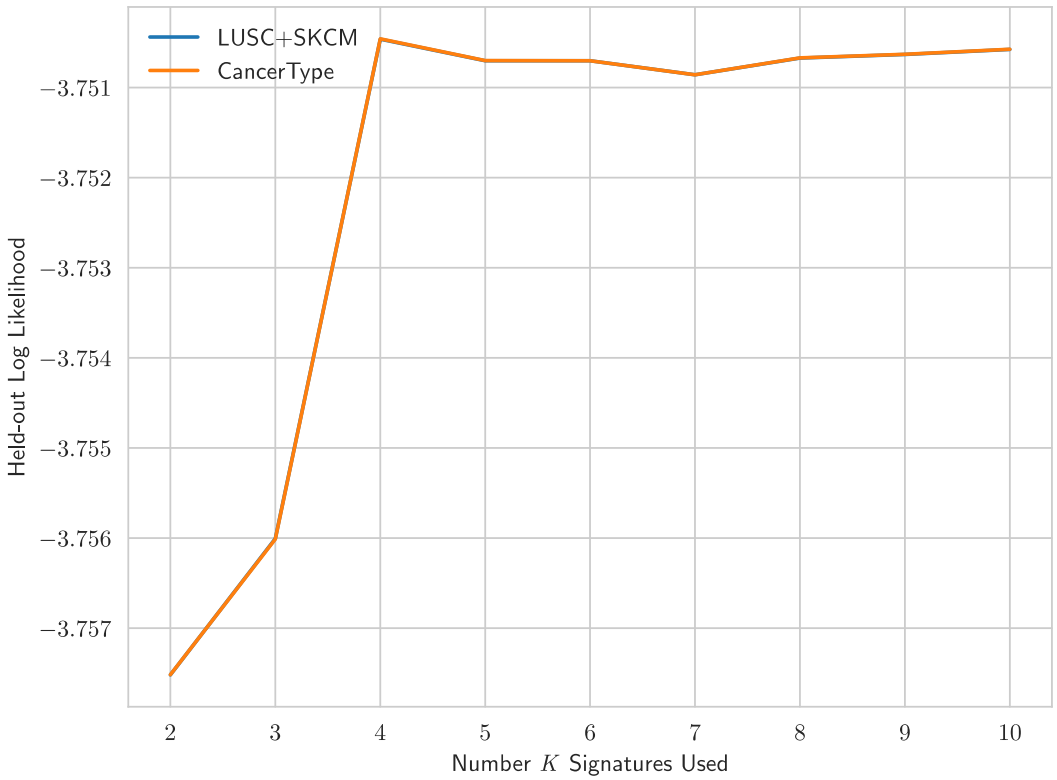


Appendix A Figure 2: The cosine similarity of the five signatures extracted TCSM on the breast cancer dataset with the biallelic HR covariate (BRCA1, BRCA2, and RAD51C) to the COSMIC signatures.

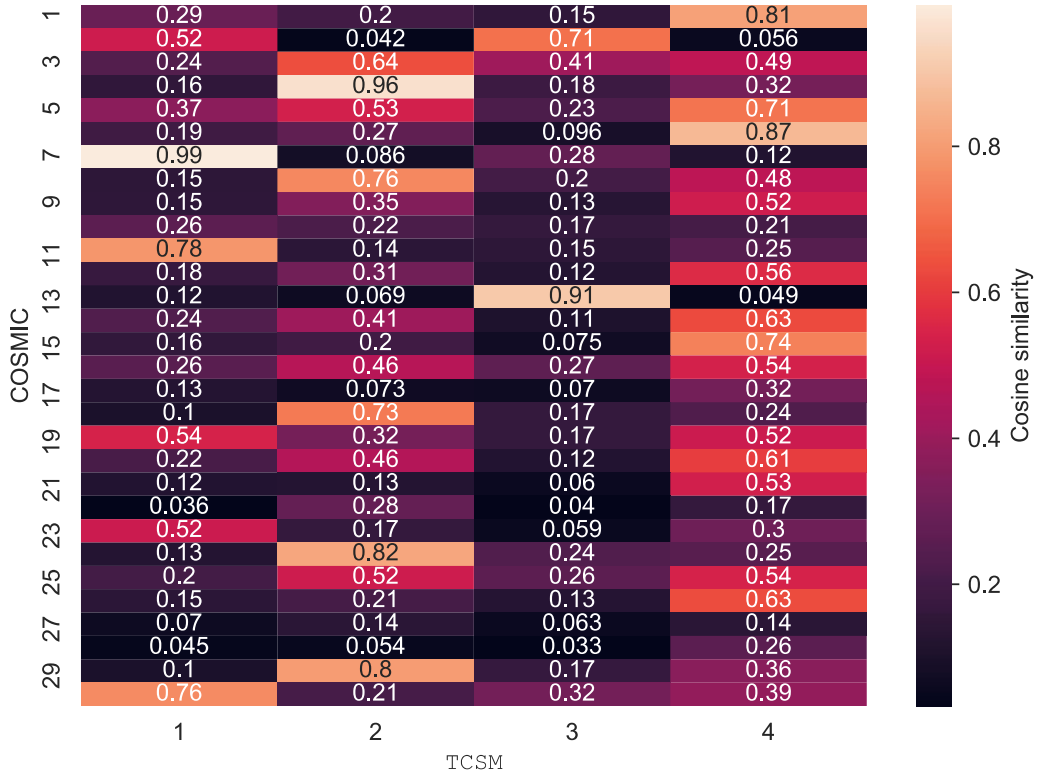
Appendix A Figure 3: The five signatures extracted by TCSM on the breast cancer dataset with the biallelic HR covariate (BRCA1, BRCA2, and RAD51C).



Appendix A Figure 4: The model selection output for the SomaticSignatures package[83] on the breast cancer dataset where the red dot is the mean value across ten runs and the crosses represent the results of individual runs.



Appendix A Figure 5: The log-likelihood performance on held-out tumors of TCSM using a single covariate to represent cancer type (Cancer Type) and TCSM using two binary covariates to represent cancer type (SKCM+LUSC)



Appendix A Figure 6: The cosine similarity of the four signatures extracted TCSM on the combined LUSC-SKCM dataset using a single covariate to represent cancer type to the COSMIC signatures.

Appendix A Figure 7: The four signatures extracted TCSM on the combined LUSC-SKCM dataset using a single covariate to represent cancer type.

Appendix B

Comparison of Salmonella-exposed cells between sequencing plates

Aulicino *et al.*[109] sequenced the infected, bystander and control cells across four sequencing plates. To understand the importance of controlling for the sequencing plate (step 2 in CSI-Microbes), we generated all 12 possible datasets with the infected cells from one plate and the exposed cells from another. Next, we ran CSI-Microbes to identify differentially abundant microbes between the infected and bystander cells (without controlling for plate) and it reported at least one genus other than *Salmonella* in nine of the twelve plate combinations.

Direct Mapping to Salmonella genomes using SRPRISM

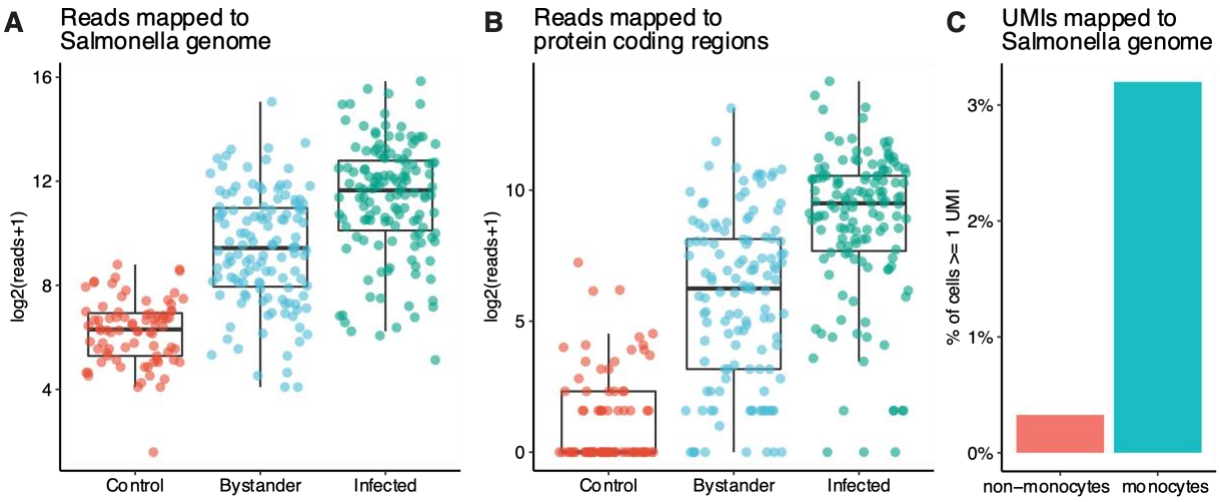
We mapped the non-human reads from Aulicino *et al.*[109] against the genomes of the *Salmonella* strains used in each study applying SRPRISM[115] to estimate the number of *Salmonella* reads present in the dataset (RefSeq release 201)[129]. We found reads that mapped to the genome of the respective *Salmonella* strains in all the infected and bystander cells as well as the control cells. Similar to the results from CSI-Microbes, we found significantly more mapped reads in the infected cells compared to both the bystander (two-sided Wilcoxon rank sum test p-value=$2e^{-11}$) and other, mock-infected cells (p-value<$2e^{-16}$) (**Appendix B Figure 1A**). Of the mapped bacterial reads, ~78% mapped to regions encoding rRNAs (including ~9% that mapped to the 7 16S rRNA genes present in each strain), while ~20% mapped to regions encoding proteins (Appendix B Direct Mapping Approach). The differential abundance between infected and bystander cells patterns was observed more strongly for reads mapped to encoding proteins (p-value=$6e^{-14}$) (**Appendix B Figure 1B**) than for reads mapped to regions encoding rRNA (p-value=$3e^{-9}$) (Appendix B Direct Mapping Approach).

Next, we applied a similar approach to the dataset from Bossel Ben-Moshe *et al.*[110] From the exposed cells, we identified 351 reads that mapped to the *SL1344* genome but the vast majority of these reads either lacked a valid cell barcode or had cell barcodes excluded by the original authors. In total, we identified 17 unique molecular identifiers (UMIs) from 15 cells expressing intracellular *Salmonella*. These cells were enriched in monocytes compared to non-monocytes (Fisher's exact test p-

value=.002, odds ration=9.75) (**Appendix B Figure 1C**) in keeping with the experimental findings of Bossel Ben-Moshe *et al.*[110]

Identification of host-transcriptomic changes associated with intracellular Salmonella

We analyzed whether we identify host cell transcriptomic changes associated with *Salmonella* infection in the cells from Aulicino *et al.* [109]. We find 318 human genes whose expression is significantly correlated with the abundance of the *Salmonella* genera (Spearman rank correlation FDR < .05). Repeating this analysis using the abundance of the other thirty-three most abundant bacterial genera, we find that the abundance of only three bacteria genera is correlated with the expression of a small number of human genes that mostly encode human ribosomal proteins. Reassuringly, we observed a strong correlation between the human genes ranked by their differential expression between the infected and bystander cells and ranked by their correlation with *Salmonella* abundance (Spearman rank correlation rho=.59, p-value=0). A gene set enrichment analysis (GSEA) of the human genes ranked by their correlation with *Salmonella* abundance identified *antigen processing and presentation of endogenous antigen* to be the most strongly down-regulated by infection by *Salmonella* (FDR q-value=$7e^4$, NES=-2.25), which mirrors the original findings of Aulicino *et al.*[109] Notably, we find that the expression of the antigen presentation genes *CD83* and *CD40* (already noted by the original authors) is negatively correlated with the expression of the *Salmonella str. D23580* acid shock protein gene *D5R57_RS08090* (Spearman correlation rho=-26, FDR=.02 (*CD40*); rho=-.28, FDR=.009 (*CD83*))[138].



Appendix B Figure 1: ***(A)*** *Number of reads per cell mapped to the Salmonella D23580 strain genome grouped by cell status from Aulicino2018.* ***(B)*** *Number of reads per cell mapped to the protein coding regions of the Salmonella D23580 strain genome grouped by cell status from Aulicino2018.* ***(C)*** *The percentage of cells with >= 1 UMI mapped to the Salmonella SL1344 strain genome grouped by cell-type from Ben-Moshe2019.*

Direct Mapping Approach

We used SRPRISM (https://github.com/ncbi/SRPRISM) with the default parameters[115] to map the unaligned reads from Bossel Ben-Moshe *et al.*[110] against the genome of *Salmonella enterica subspecies enterica serovar Typhimurium strain SL1344* (RefSeq assembly accession: GCF_000210855.2)[139] and the unaligned reads from Aulicino *et al.*[109] against the *Salmonella enterica subspecies enterica serovar Typhimurium strain LT2* (RefSeq assembly accession: GCF_000006945.2) [140] and *strain D23580* (RefSeq assembly accession: GCF_900538085.1)[141] because it is more tolerant of errors than other more commonly used read alignment tools. We assigned reads to genes for the Smart-seq2 dataset using the intersect command of bedtools[142] and the gene feature format (GFF) file associated with the RefSeq assembly accession. We used the attributes column from the GFF file to assign reads as either

protein coding (gene_biotype=protein_coding) or encoding ribosomal RNA (gene_biotype=rRNA).

Gene Set Enrichment Analysis

We performed gene set enrichment analysis using the GSEAPreranked function of the GSEA tool v4.1.0 with default settings and seed=149[134] with the gene ontology (GO) biological processes v7.3 gene set from MSigDB[134–137]. To perform GSEA on the *Salmonella* dataset, we first calculated the Spearman rank correlation between the spike-in normalized abundance of the *Salmonella* genus and the spike-in normalized expression of human genes expressed above 10 counts per million (CPM) in at least 50% of the cells using the correlatePairs function from scran[131]. Next, we ranked genes by multiplying the $-\log_{10}$(p-value) by -1 (Spearman rank correlation rho >0) or 1 (rho <=0).

Bibliography

1. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, (2008).
2. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, (2005).
3. Wetterstrand, K. A. The Cost of Sequencing a Human Genome. *National Human Genome Research Institute* https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost (2020).
4. Miller, J. R., Koren, S. & Sutton, G. Assembly algorithms for next-generation sequencing data. *Genomics* vol. 95 (2010).
5. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, (2009).
6. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
7. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. **00**, 1–3 (2013).
8. Burrows, M. & Wheeler, D. J. A Block-sorting Lossless Data Compression Algorithm. **28**, 409–422 (1979).
9. Hodges, E. *et al.* Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.* **39**, (2007).
10. Koboldt, D. C. *et al.* VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, (2012).

11. Koboldt, D. C. *et al.* VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**, (2009).

12. Van der Auwera, G. A. *et al.* From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* (2013) doi:10.1002/0471250953.bi1110s43.

13. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics* vol. 10 (2009).

14. Garraway, L. A. & Lander, E. S. Lessons from the cancer genome. *Cell* vol. 153 (2013).

15. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, (2009).

16. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, (2011).

17. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).

18. Brennan, C. W. *et al.* The somatic genomic landscape of glioblastoma. *Cell* **155**, (2013).

19. McLendon, R. *et al.* Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, (2008).

20. Koboldt, D. C. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490**, (2012).

21. Hudson, T. J. *et al.* International network of cancer genome projects. *Nature* vol. 464 (2010).

22. Quesada, V. *et al.* Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat. Genet.* **44**, (2012).

23. Chapman, M. A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, (2011).

24. Dees, N. D. *et al.* MuSiC: Identifying mutational significance in cancer genomes. *Genome Res.* **22**, (2012).

25. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, (2013).

26. Leiserson, M. D. M., Wu, H., Vandin, F. & Raphael, B. J. CoMEt : a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome Biol.* 1–20 (2015) doi:10.1186/s13059-015-0700-7.

27. Leiserson, M. D. M. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47**, (2015).

28. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–21 (2013).

29. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).

30. Farmer, H. *et al.* Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* **434**, 917–921 (2005).

31. Riaz, N. *et al.* Pan-cancer analysis of bi-allelic alterations in homologous recombination DNA repair genes. *Nat. Commun.* 1–7 (2017)

doi:10.1038/s41467-017-00921-w.

32. Pfeifer, G. P. *et al.* Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene* **21**–**48**, 7435–7451 (2002).

33. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Rep.* **3**, 246–259 (2013).

34. Shiraishi, Y., Tremmel, G., Miyano, S. & Stephens, M. A Simple Model-Based Approach to Inferring and Visualizing Cancer Mutation Signatures. *PLoS Genet.* **11**, 1–21 (2015).

35. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).

36. Blei, D. M. & Lafferty, J. D. A correlated topic model of Science. *Ann. Appl. Stat.* **1**, 17–35 (2007).

37. Ramage, D., Hall, D., Nallapati, R. & Manning, C. D. Labeled LDA : A supervised topic model for credit attribution in multi-labeled corpora. *Proc. Conf. Empir. Methods Nat. Lang. Process.* 248–256 (2009).

38. Mimno, D. & McCallum, A. Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression. in *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence* 411–418 (2008). doi:10.1.1.140.6925.

39. Funnell, T. *et al.* Integrated structural variation and point mutation signatures in cancer genomes using correlated topic models. *PLoS Comput. Biol.* **15**, 1–24 (2019).

40. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).

41. Robinson, W., Sharan, R. & Leiserson, M. D. M. Modeling clinical and molecular covariates of mutational process activity in cancer. *Bioinformatics* **35**, i492–i500 (2019).

42. Rosenberg, S. A. & Restifo, N. P. Adoptive cell transfer as personalized immunotherapy for human cancer. *Science (80-.).* **348**, 62–68 (2015).

43. Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).

44. Sender, R., Fuchs, S. & Milo, R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biol.* **14**, 1–14 (2016).

45. Hiergeist, A., Gläsner, J., Reischl, U. & Gessner, A. Analyses of intestinal microbiota: Culture versus sequencing. *ILAR J.* **56**, 228–240 (2015).

46. Sepich-Poore, G. D. *et al.* The microbiome and human cancer. *Science* **371**, (2021).

47. Bullman, S. *et al.* Analysis of Fusobacterium persistence and antibiotic response in colorectal cancer. *Science (80-.).* **358**, 1443–1448 (2017).

48. Castellarin, M. *et al.* Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome Res.* **22**, 299–306 (2012).

49. 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).

50. Kostic, A. D. *et al.* PathSeq: A comprehensive computational tool... *Nat.*

Biotechnol. **29**, 393–396 (2011).

51. Wood, D. E. & Salzberg, S. L. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, (2014).

52. 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).

53. Kostic, A. D. *et al.* Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Res.* **22**, 292–298 (2012).

54. 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).

55. 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).

56. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12**, 1–12 (2014).

57. Dohlman, A. B. *et al.* The cancer microbiome atlas: a pan-cancer comparative analysis to distinguish tissue-resident microbiota from contaminants. *Cell Host Microbe* **29**, 281-298.e5 (2021).

58. Poore, G. D. *et al.* Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* **579**, 567–574 (2020).

59. Tomasetti, C; Li, L; Vogelstein, B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science (80-.).* **355**, 1330–1334

(2017).

60. Tubbs, A. & Nussenzweig, A. Endogenous DNA Damage as a Source of Genomic Instability in Cancer. *Cell* **168**, 644–656 (2017).

61. Le, T. D., Durham, N., Smith, N., Wang, H. & Bartlett, B. Mismatch-repair deficiency predicts response of solid tumors to PD-1 blockade. *Science (80-.).* **6733**, (2017).

62. Hegi, M. E. *et al.* MGMT gene silencing and benefit from temozolomide in glioblastoma. *N. Engl. J. Med.* **352**, 997–1003 (2005).

63. Davies, H. *et al.* HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.* **23**, 517–525 (2017).

64. Pfeifer, G. P. Environmental exposures and mutational patterns of cancer genomes. *Genome Med.* **2**, 2–5 (2010).

65. Szikriszt, B. *et al.* A comprehensive survey of the mutagenic impact of common cancer cytotoxics. *Genome Biol.* **17**, 1–16 (2016).

66. Trucco, L. D. *et al.* Ultraviolet radiation–induced DNA damage is prognostic for outcome in melanoma. *Nat. Med.* (2018) doi:10.1038/s41591-018-0265-6.

67. Chawanthayatham, S. *et al.* Mutational spectra of aflatoxin B 1 in vivo establish biomarkers of exposure for human hepatocellular carcinoma. *Proc. Natl. Acad. Sci.* **114**, E3101–E3109 (2017).

68. Forbes, S. A. *et al.* COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).

69. Kasar, S. *et al.* Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia

evolution. *Nat. Commun.* **6**, 8866 (2015).

70. Fischer, A., Illingworth, C. J. R., Campbell, P. J. & Mustonen, V. EMu: Probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biol.* **14**, (2013).

71. Rosales, R. A., Drummond, R. D., Valieris, R., Dias-Neto, E. & Da Silva, I. T. signeR: An empirical Bayesian approach to mutational signature discovery. *Bioinformatics* **33**, 8–16 (2017).

72. Huang, X., Wojtowicz, D. & Przytycka, T. M. Detecting presence of mutational signatures in cancer with confidence. *Bioinformatics* **34**, 330–337 (2018).

73. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. deconstructSigs: Delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 1–11 (2016).

74. Blei, D., Carin, L. & Dunson, D. Probabilistic topic models. *IEEE Signal Process. Mag.* **27**, 55–65 (2012).

75. Alexandrov, L. B. *et al.* Mutational signatures associated with tobacco smoking in human cancer. *Science (80-.).* **354**, 618–622 (2016).

76. Polak, P. *et al.* A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat. Genet.* (2017) doi:10.1038/ng.3934.

77. Kim, J. *et al.* Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.* **48**, 600–606 (2016).

78. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).

79. Roberts, M. E., Stewart, B. M., Tingley, D. & Airoldi, E. M. The structural topic model and applied social science. *NIPS 2013 Work. Top. Model.* 2–5 (2013).

80. Roberts, M. E., Stewart, B. M. & Tingley, D. Navigating the local modes of big data: The case of topic models. *Comput. Soc. Sci. Discov. Predict.* 51–97 (2016) doi:10.1017/9781316257340.004.

81. Wallach, H. M., Murray, I., Salakhutdinov, R. & Mimno, D. Evaluation methods for topic models. *Proc. 26th Annu. Int. Conf. Mach. Learn. - ICML '09* 1–8 (2009) doi:10.1145/1553374.1553515.

82. 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).

83. Gehring, J. S., Fischer, B., Lawrence, M. & Huber, W. SomaticSignatures: Inferring mutational signatures from single-nucleotide variants. *Bioinformatics* **31**, 3673–3675 (2015).

84. Köster, J. & Rahmann, S. Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics* **28**, 2520–2522 (2012).

85. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291-304.e6 (2018).

86. Rieunier, G., Caux-moncoutier, V., Tirapo, C., Popova, T. & Mani, E. Ploidy

and Large-Scale Genomic Instability Consistently Identify Basal-like Breast Carcinomas with BRCA1 / 2 Inactivation. **72**, 5454–5463 (2012).

87. Knijnenburg, T. A. *et al.* Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. *Cell Rep.* **23**, 239–254 (2018).

88. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).

89. Vandin, F., Upfal, E. & Raphael, B. J. De novo discovery of mutated driver pathways in cancer. 375–385 (2012) doi:10.1101/gr.120477.111.22.

90. Miller, J. H. Mutagenic specificity of ultraviolet light. *J. Mol. Biol.* **182**, 45–65 (1985).

91. Campbell, J. D. *et al.* Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat. Genet.* **48**, 607–616 (2016).

92. Campbell, J. D. *et al.* Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat. Genet.* **48**, 607–616 (2016).

93. Teh, Y. W., Jordan, M. I., Beal, M. J. & Blei, D. M. Hierarchical Dirichlet Processes. 1–30 (2005).

94. Eisenstein, J., Ahmed, A. & Xing, E. P. E. Sparse additive generative models of text. *Proc. 28th Int. Conf. Mach. Learn.* 1041–1048 (2011) doi:10.1.1.206.5167.

95. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends*

Genet. **29**, 569–574 (2013).

96. Robinson, W., Schischlik, F., Gertz, E. M., Schäffer, A. A. & Ruppin, E. Identifying the Landscape of Intratumoral Microbes via a Single Cell Transcriptomic Analysis. *bioRxiv* 1–11 (2020).

97. Gur, C. *et al.* Fusobacterium nucleatum supresses anti-tumor immunity by activating CEACAM1. *Oncoimmunology* **8**, e1581531 (2019).

98. Yu, T. C. *et al.* Fusobacterium nucleatum Promotes Chemoresistance to Colorectal Cancer by Modulating Autophagy. *Cell* **170**, 548-563.e16 (2017).

99. Pleguezuelos-Manzano, C. *et al.* Mutational signature in colorectal cancer caused by genotoxic pks + E. coli. *Nature* **580**, 269–273 (2020).

100. 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).

101. 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).

102. Nejman, D. *et al.* The human tumor microbiome is composed of tumor type-specific intracellular bacteria. *Science (80-.).* **368**, 973–980 (2020).

103. Kalaora, S. *et al.* Identification of bacteria-derived HLA-bound peptides in melanoma. *Nature* (2021) doi:10.1038/s41586-021-03368-8.

104. 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).

105. 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).

106. Steuerman, Y. *et al.* Dissection of Influenza Infection In Vivo by Single-Cell RNA Sequencing. *Cell Syst.* **6**, 679-691.e4 (2018).

107. Bost, P. *et al.* Host-Viral Infection Maps Reveal Signatures of Severe COVID-19 Patients. *Cell* **181**, 1475-1488.e12 (2020).

108. 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).

109. Aulicino, A. *et al.* Invasive Salmonella exploits divergent immune evasion strategies in infected and bystander dendritic cell subsets. *Nat. Commun.* **9**, 4883 (2018).

110. 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).

111. 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).

112. Stouffer, S. A., Suchman, E. A., DeVinney, L. C. & Williams, R. M. *The American Soldier. Adjustment During Army LIfe*. (Princeton University Press, 1949).

113. Yekutieli, D. Hierarchical false discovery rate-controlling methodology. *J. Am. Stat. Assoc.* **103**, 309–316 (2008).

114. Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Res.* **40**, D136-143 (2012).

115. Morgulis, A. & Agarwala, R. SRPRISM (Single Read Paired Read Indel Substitution Minimizer): an efficient aligner for assemblies with explicit guarantees. *Gigascience* **9**, (2020).

116. Paulson, K. G. *et al.* Acquired cancer resistance to combination immunotherapy from transcriptional loss of class I HLA. *Nat. Commun.* **9**, (2018).

117. Lee, H. O. *et al.* Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nat. Genet.* **52**, 594–603 (2020).

118. 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).

119. Maynard, A. *et al.* Therapy-Induced Evolution of Human Lung Cancer Revealed by Single-Cell RNA Sequencing. *Cell* **182**, 1232-1251.e22 (2020).

120. Eishi, Y. Etiologic link between sarcoidosis and Propionibacterium acnes. *Respir. Investig.* **51**, 56–68 (2013).

121. 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).

122. 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).

123. 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).

124. Radhakrishnan, G. K. & Splitter, G. A. Modulation of host microtubule dynamics by pathogenic bacteria. *Biomol. Concepts* **3**, 571–580 (2012).

125. Mohanty, B. K. & Kushner, S. R. Bacterial/archaeal/organellar polyadenylation. *Wiley Interdiscip Rev RNA* **2**, 256–276 (2010).

126. 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).

127. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

128. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

129. 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).

130. Schäffer, A. A. *et al.* VecScreen-plus-taxonomy: Imposing a tax(onomy) increase on vector contamination screening. *Bioinformatics* **34**, 755–759 (2018).

131. 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).

132. Sokal, R. R. & Rohlf, F. J. *Biometry*. (W. H. Freeman, 1995).

133. Sankaran, K. & Holmes, S. structSSI: Simultaneous and selective inference for grouped or hierarchically structured data. *J. Stat. Softw.* **59**, 1–21 (2014).

134. 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).

135. Ashburner, M. *et al.* Gene ontology: Tool for the unification of biology. *Nature Genetics* vol. 25 (2000).

136. Carbon, S. *et al.* The Gene Ontology resource: Enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334 (2021).

137. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).

138. Kretschmer, B., Kühl, S., Fleischer, B. & Breloer, M. Activated T cells induce rapid CD83 expression on B cells by engagement of CD40. *Immunol. Lett.* **136**, 221–227 (2011).

139. Kröger, C. *et al.* The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 1277–1286 (2012).

140. McClelland, M. *et al.* Complete genome sequence of Salmonella enterica serovar Typhimurium LT2. *Nature* **413**, (2001).

141. Canals, R. *et al.* Adding function to the genome of African Salmonella Typhimurium ST313 strain D23580. *PLoS Biol.* **17**, e3000059 (2019).

142. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).