# BacTIME: Computational inference of bacterial interactions with the tumor microenvironment

**Anonymous Authors**[1]

## Abstract

Bacteria inside of tumors can be innately therapeutic through competition for nutrients and by stimulating a local immune response, spurring efforts to engineer bacteria as anti-cancer therapies. Multiple computational approaches have been developed for inferring interactions between tumor and immune cells from single-cell and spatial genomic data, yet we lack analytical frameworks by which to infer the consequence of tumor-bacterial interactions. This study extends a deep generative framework to dissect tumor-bacterial interactions and resolve microbial communities from spatial transcriptomic data. Our inferred latent representation disentangles the bacterial interactions with tumor-immune microenvironment (TIME) populations in a way that pure spatial representation would not be able to. Our results revealed distinct intratumoral immune subpopulations characterized by their interaction with bacteria, including a subset of dysfunctional and activated T cells, and a trajectory of bacteria-associated monocytes characterized by upregulation of innate signaling pathways. These findings highlight the potential of integrative analysis of spatial and single-cell transcriptomic data with taxonomically aligned inference of tumor-associated bacteria to analyze microbial communities in cancer.

## 1. Introduction

Recently, studies have suggested that microbes can significantly impact the development, progression, and response to cancer therapy [1, 2, 3]. For example, studies that probe microbiome interactions with colorectal cancer cells identified shifts in immunologic and metastatic molecular states [2, 4]. However, the underlying mechanisms linking microbes and changes in tumor state remain poorly understood.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by *the 2023 ICML Workshop on Computational Biology*. Do not distribute.

Studying the interaction between bacteria and the TIME has largely relied on metagenomic sequencing to profile microbial communities and fluorescence in situ hybridization (FISH) or immunohistochemistry to characterize the spatial distribution of bacteria. Spatial and single-cell transcriptomics techniques offer greater depth and higher resolution tools to study the host tissue. Still, we are limited in studying the microbial transcriptome due to the eukaryotic-specific chemistry leveraged for mRNA capture [5]. This makes it challenging to infer bacterial-host interactions from transcriptomic data alone. However, it is possible that mispriming events during transcriptomic library preparation may capture some bacterial rRNA, which then could be aligned to multiple bacteria genomes [6]. While this technique has been implemented to highlight the localization of bacteria to certain tissue areas, there is a dearth of computational methods that integrate this additional "mode" of bacterial transcripts. Thus, there is a strong demand for computational methods to utilize this data in modeling bacterial-TIME interactions, while accounting for the sparsity of transcripts and spurious alignments to any one of the thousands of possible microbe references.

Spatial deconvolution algorithms face additional challenges with the integration of bacterial data, as they often rely on single-cell atlases or prior knowledge of single organism cell-states and interactions. For the purpose of integrating bacterial reads, we do not have prior knowledge of which host cells interact with bacteria spatially or how any populations of cells change phenotypically in the presence of different bacteria, which demands a novel analytical framework. In particular, constructing informative priors on bacterial-associated regions is challenging. Deconvolving bacteria and TIME cells in the same spots (i.e., spatial location) is thus not a straightforward task. Finally, cell-cell interaction inference techniques, such as CellphoneDB [7] or Cellchat [8], do not take non-host cells into account for learning interactions, because the databases employed are curated using only host receptor-ligand pairs.

We propose to address the numerous challenges presented by adapting a recently developed deep generative model, Starfysh [9] to learn and deconvolve information on bacterial-TIME interactions. The novelty of our approach is in adapting the prior to disentangle bacteria-associated

spots, using archetypal analysis [9]. We demonstrate this approach allows use of existing cell-type marker information, bacterial genome references and histology images to learn the spatial organization of bacteria as additional cell-types and cell-states in the TIME interacting with bacteria. We then validate and extend the findings from this model in a parallel single-cell RNA-seq (scRNA) dataset, as an independent measurement from the same system.

# 2. Model

## 2.1. Datasets

Raw fastq sequencing files were obtained after performing either 10x Visium spatial transcriptomic (ST) or 10x Chromium scRNA-seq on tumors harvested from BALB/c mice with subcutaneous CT26 colorectal cancer tumors. For the Visium dataset, tumors were harvested 10 days after systemic delivery of *E. coli Nislux* (*E. coli Nissle* with an engineered Lux cassette that allows for easier visualization of bacteria [10]), through tail vein injection. For the scRNA-seq dataset, tumors were harvested 10 days after intratumoral injection of either control PBS or *E. coli Nislux*. By leveraging a dataset with a controlled *in vivo* delivery of a known bacterial strain, we can experimentally validate the ability of our pipeline in distinguishing *E. coli* reads from other taxonomies.

## 2.2. Construction of the bacteria prior

To construct a prior for bacteria-associated spots in ST data, we first leveraged the GATK Pathseq pipeline [11] to identify sequencing reads unaligned to the host organism that align to a reference compendium containing microbial genomes. While we do not expect 10x Visium or 10x Chromium scRNA-seq 3' end capture to target microbial RNA (lacking a poly-A tail), polyT primers used for eukaryotic messenger RNA capture display a rate of mispriming that captures a subset of microbial reads. In brief, taxa were assigned to reads with a minimum clip length of 60 bp, filter-duplicates were set to false to avoid loss of duplicate reads, and the identity-score was set to 0.7 as performed previously [12]. The output from the GATK Pathseq pipeline was a tagged list of microbial reads which was then processed with the Python package Pysam. For the spatial analysis, we integrated the bacterial reads with the mouse gene expression matrix by appending the bacteria as features in the expression matrix. For the consequent single cell analysis, the bacterial reads are included as metadata.

## 2.3. Starfysh and adaptation

The Starfysh algorithm [9] takes as inputs a spatial transcriptomic (ST) dataset, optional signature gene lists for cell types or cell states, and an optional paired histology image. To infer cell proportions and densities, the algorithm compresses transcriptomic data ($x_i$ for spot $i$) and guides the low dimensional representation ($z_i$) with *anchors* defined as spots enriched for known markers of cell types and states, which form priors in the deep generative model. Starfysh transforms the latent variable $z_i$ via the neural network $f$ followed by scaling with $l_i$, which is sampled from a log-normal distribution according to observed library size while also accounting for spatial dependencies in cell density between adjacent spots [9]. The model assumes that the histology variable ($y_i$) is jointly generated from the latent embedding ($z_i$).

Extending this model, we assume the observed gene transcripts ($x_{ig}$ for gene $g$ and spot $i$, and $x_{im}$ for bacteria genus $m$ and spot $i$) are sampled from a negative binomial distribution where

$$p_\theta(x_{ig}|l_i, z_i) = NegativeBinomial(l_i \cdot f(z_i), \theta_g)$$

where $\theta_g$ represents gene-specific dispersions. The prior for cell type proportions $c_{ik}$ for cell type $k$ is computed according to $A(x_i, s_k)$ with $s_k$ denoting the gene expression signature for cell type $k$ if $i$ is an anchor spot and 0 otherwise, thus relying on anchors to deconvolve data.

$$p(c_i, \alpha) = Dirichlet(\alpha \cdot A(x_i, s))$$
$$A(x_i, s_k) = (\frac{\sum_{g\in[1,\dots G]} x_{ig} \cdot s_{kg}}{\sum_g s_{kg}} - u_k) \cdot \frac{1}{\sigma_k}$$

while $\alpha$ can tune the strength of the prior. Finally, the generative model integrates all the above, and assumes that $z_i$ is normally distributed.

$$p(l_i, z_i, x_i, c_i) = p(x_i|z_i, l_i)p(c_i, \alpha)p(z_i)p(l_i, \tilde{l_l})$$

The novelty of our work in extending the archetypal analysis feature of Starfysh to identify additional anchors corresponding to bacteria-associated spots. Intuitively, archetypal analysis fits a convex polytope to the observed data, finding the prototypes (archetypes) that are most adjacent to the extrema of the data manifold. We hypothesize that introducing bacteria into tissues in a controlled setting would lead to new observed cell states based on bacteria-TIME interactions. Therefore, in order to characterize these new cell states and interactions that have not been previously studied, we reasoned that applying archetypal analysis is appropriate as the TIME changes upon perturbation through bacterial delivery should be represented by at least one extrema in the manifold with dimensions spanning both host and bacteria genes.

We thus define expanded variables for counts per spot as $x_i = [x^T_{i,1,\cdots,G}, x^T_{i,1,\cdots,M}]^T$ with $G$ total genes and $M$ bacteria genera. We also define a new signature $s_b$ for microbial-related effects and expand proportions as $c_i = [c^T_{i,1,\cdots,G}, c_{i,b}]^T$, where $c_b$ summarizes signatures $s_m$ corresponding to the *E. coli-Shigella* genera $m \in [1, \dots M]$ :

$$A(x_{im}, s_b) = (\frac{\sum_{m\in[1,\dots,M]} x_{im} \cdot s_{bm}}{\sum_m s_{bm}} - u_b) \cdot \frac{1}{\sigma_b}$$

Finally, $p(c_{ib}, \alpha) = Dirichlet(\alpha \cdot A(x_{im}, s_b))$

We expect that the addition of bacterial transcriptome information with the gene level information will empower Starfysh to characterize a novel context-specific cell state that relates to association of TIME cells with bacteria. After learning a bacteria-associated archetype, this cell state will be incorporated as a prior to the model and further refine the deconvolution of cell types with the model.

## 3. Results and discussion

### 3.1. Spatial archetypal analysis

We applied the adapted model to dissect bacteria-TIME interaction in the mouse Visium ST dataset. Our inferred cell type proportions reveal a tumor region with increased CT26 tumor proportion in the top portion of the histology image (Fig. 1a), which overlaps with both the regulatory T cell (Treg) and monocyte-rich areas of the tumor and also colocalizes with regions where precursor exhausted T cells are enriched (Fig. 1b, g, f). In this same region, terminally exhausted T cells are found at relatively lower proportions compared to their inferred proportions across the rest of the tissue (Fig. 1c). Strikingly, the deconvolved bacterial population is also geographically found in regions of increased CT26 proportions. This observation is consistent with previous studies that use engineered bacteria to treat cancer [13, 14, 15]. Therefore, as we hypothesized, systemically delivered microbes travel and preferentially grow in immunocompromised and hypoxic neighborhoods of increased tumor cell activity. However, we notice that archetypal analysis identifies only a subset of the bacteria-enriched as a novel phenotypic state (Fig. 1d, h).

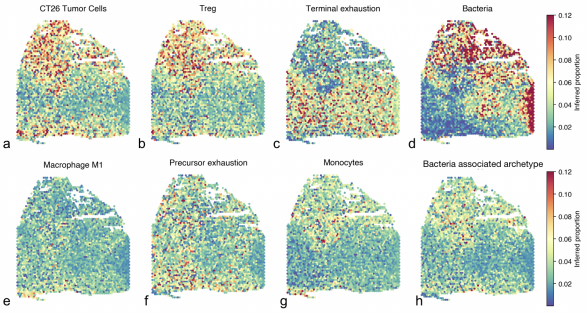


*Figure 1.* Inferred cell type proportions in the spatial map corresponding to the paired histology image

We next clustered spots in the $c$-space by their inferred cell type proportions identifying regions with unique cell type composition (Fig. 2a, b). Generally, the spots with the highest proportion of each cell type group together along vertices of the UMAP embedding. In particular, clusters 1 and 2 capture the bacteria-associated archetype and its heterogeneity among spots (Fig. 2b, c). This shows that the inferred proportions are interpretable among all cell types considered, including the bacteria-associated archetype. We found that cluster 1 is mainly composed of Tregs, the bacteria-associated archetype, bacteria, CT26, and monocytes, while cluster 2 includes the bacteria-associated archetype, CT26, monocytes, M1 macrophages, precursor exhaustion T cells, and dysfunctional T cells (Fig. 2d). Interestingly, cluster 2 suggests colocalization and possible communication between a subset of bacteria and dysfunctional T cell states.

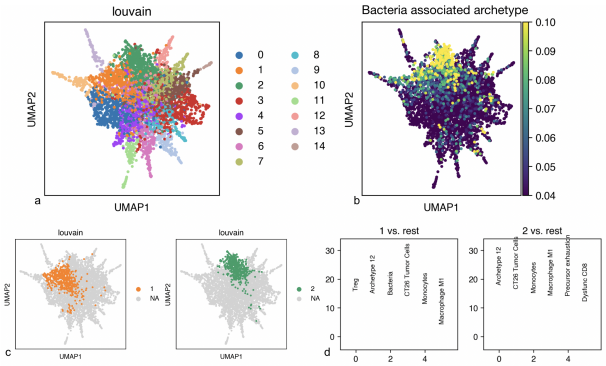


*Figure 2.* UMAP of the $c$ space colored by a) Louvain clusters, b) the inferred bacteria-associated archetype proportion, c) clusters 1 and 2; d) Inferred cluster specific cell types for clusters 1 and 2

We performed diffusion component (DC) analysis on the inferred latent z-space of our model, to identify trajectories of cell state transitions. We observed that the spots with high bacteria proportion are focused primarily on the rightmost branch along DC3 (Fig. 3h). The spots enriched in CT26, Treg, and monocytes overlap with the bottom region of the bacteria-rich spots (Fig. 3a, b, g, h). Interestingly, that same bacteria region colocalizes with the highest proportion of bacteria-associated archetype, dysfunctional CD8 cells, precursor and terminally exhausted T cells, while the top of this bacteria-rich region colocalizes with spots of high activated CD8 proportion (Fig. 3c-h). We further quantified these observations by calculating the Pearson product-moment correlation between the inferred cell type proportions per spot, DC2 and DC3 (Fig. 3j), showing a positive correlation between bacteria, activated CD8 and DC2, and a negative correlation between bacteria-associated archetype and the dysfunctional T cell state. These results highlight the interpretability of our model's latent space, as it allows us to identify trajectories that lead to the emergence of two distinct bacteria populations: a branch correlated with bacteria and T cell activation, and a second with bacteria and T cell dysfunction. Further analysis can pinpoint mechanisms driving this bifurcation.

Therefore, by adapting the priors in Starfysh and interpreting the inferred latent representation, we were able to disentangle two distinct bacteria populations associated with different immune responses in the TIME, which cannot be discerned solely from the spatial representation of our dataset.

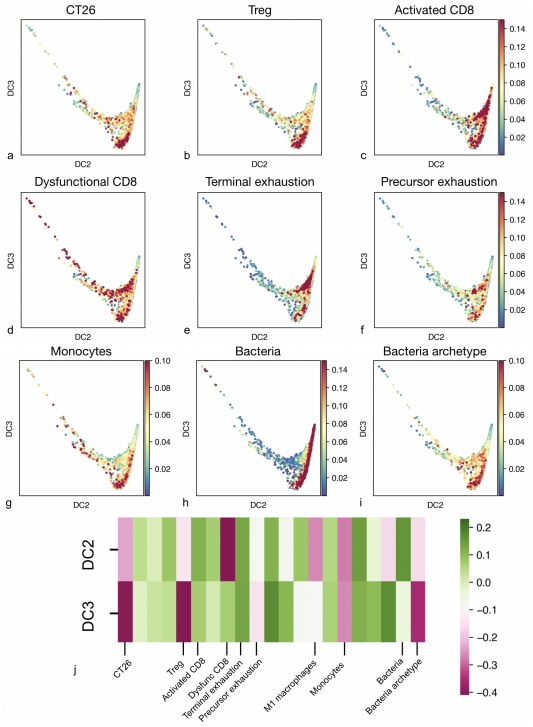


*Figure 3.* a-i Diffusion map projections along DC2 and DC3 of the Visium spots colored by inferred proportion of 9 individual cell types; j) Correlation coefficients between DC2, DC3 and inferred proportions of all cell types considered in the gene signature list

### 3.2. Single-cell analysis

Following the archetypal analysis predicting association of dysfunctional T cells, monocytes, and the novel bacteria-associated archetype, we sought to determine if a similar pattern was present in single-cell resolution data using matched scRNA data. Indeed we observe a high association of *E. coli* reads specific to the monocyte clusters (Appendix Fig. 4). Interestingly, only monocyte clusters showed high association with bacterial reads, whereas macrophage clusters did not show high percentage of bacteria positive cells. This is consistent with our spatial colocalization analysis which identified an association of monocytes and the bacterial archetype, and lower colocalization of macrophages. Moreover, our results agree with additional studies using RNA-FISH to detect bacterial rRNA in the tumor microenvironment, observing a low number of bacterial reads detected in macrophages compared to other immune cell types, possibly due to degradation within activated macrophages [16].

Clustering analysis identified a subpopulation of monocytes enriched for bacterial association, and subsequent DEG analysis across clusters identified the monocyte subpopulation overexpresses Wfdc21–a gene downstream of LPS (a component of bacterial cell walls) recognition. Diffusion component analysis of innate immune populations (macrophages and monocytes) identified a trajectory (DC1) describing a cell state with a high association with bacteria (Fig. 4a, b) [17]. Examining the genes highly correlated

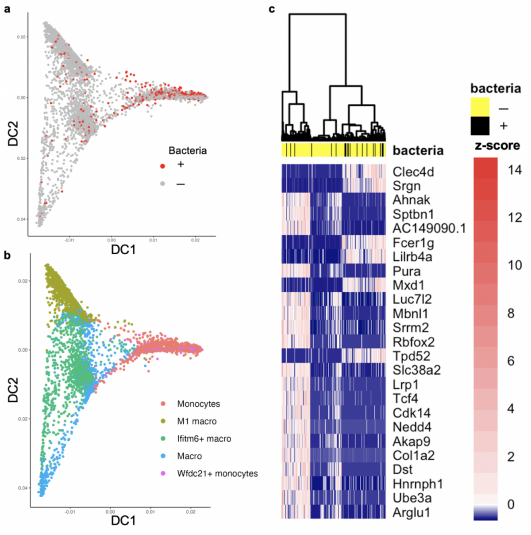


*Figure 4.* a-b Diffusion map projections along DC1 and DC2 of scRNA for merged + and -bacteria conditions c) heatmap of z-scored expression values across genes highly correlated with DC1

with DC1, we identified an association with innate immune function genes such as the Clec4d–mycobacterial receptor, Srgn–hematopoietic cell granule proteoglycan and the Lilrb4–Leukocyte immunoglobulin-like receptor involved in regulation of mast cell activation (Fig. 3c) [17]. Upregulation of these genes in the bacterial associated populations may contribute to alterations in T-cell signaling and priming. Finally, we wanted to define which genes drive formation of this trajectory while taking into account whether or not the cells are associated with bacteria. We fit a quasi-poisson regression model using coordinates across DC1 and bacteria positive status as parameters. This analysis identified Lcn2-an iron-sequestering innate immune response protein, Tnfrsf13b-a TNF receptor involved in regulating T-cell and B-cell stimulation, and Irak3-a regulator of Toll-like receptor signaling as putative regulators of innate immune progression along the trajectory of association with bacteria [17].

## 4. Conclusion

Our adaptation of the Starfysh method with the additional layer of bacterial rRNA reads marks an important step for the disentangling of bacterial-TIME interactions. The uncovered archetype by our model arrives at an easily interpretable, biologically meaningful phenotype that may impact bacterial or immune therapy. The scalability of introducing microbial rRNA into archetypal and subsequent scRNA data is also a novel feature that can inform precise engineering of bacterial immunotherapy while gaining insights into the cell states driven by microbes.

# References


[1] Chi Ma et al. "Gut microbiome-mediated bile acid metabolism regulates liver cancer via NKT cells". In: *SCIENCE* 360.6391 (May 2018).

[2] Michael A Casasanta et al. "Fusobacterium nucleatum host-cell binding and invasion induces IL-8 and CXCL1 secretion that drives colorectal cancer cell migration". In: *Science signaling* 13.641 (2020), eaba9157.

[3] Yujie Bao et al. "Long noncoding RNA BFAL1 mediates enterotoxigenic Bacteroides fragilis-related carcinogenesis in colorectal cancer via the RHEB/mTOR pathway". In: *Cell Death & Disease* 10.9 (2019), p. 675.

[4] Ce Yuan et al. "Tumor models to assess immune response and tumor-microbiome interactions in colorectal cancer". In: *Pharmacology & therapeutics* 231 (2022), p. 107981.

[5] Antoine-Emmanuel Saliba et al. "Single-cell RNA-seq: advances and future challenges". In: *Nucleic Acids Research* 42.14 (July 2014), pp. 8845–8860.

[6] Grace E Johnson et al. "BaM-seq and TBaM-seq, highly multiplexed and targeted RNA-seq protocols for rapid, low-cost library generation from bacterial samples". In: *NAR Genomics and Bioinformatics* 5.1 (2023).

[7] Mirjana Efremova et al. "CellPhoneDB v2. 0: Inferring cell-cell communication from combined expression of multi-subunit receptor-ligand complexes". In: *doi* 10 (2019), p. 680926.

[8] Suoqin Jin et al. "Inference and analysis of cell-cell communication using CellChat". In: *Nature communications* 12.1 (2021), p. 1088.

[9] Siyu He et al. "Starfysh reveals heterogeneous spatial dynamics in the breast tumor microenvironment". In: *bioRxiv* (2022).

[10] Sreyan Chowdhury et al. "Programmable bacteria induce durable tumor regression and systemic antitumor immunity". In: *Nature Medicine* (2019).

[11] Mark A Walker et al. "GATK PathSeq: a customizable computational tool for the discovery and identification of microbial sequences in libraries from eukaryotic hosts". In: *Bioinformatics* 34.24 (July 2018), pp. 4287–4289.

[12] Jorge Luis Galeano Niño et al. "Effect of the intratumoral microbiota on spatial and cellular heterogeneity in cancer". In: *Nature* 611.7937 (2022), pp. 810–817.

[13] Sreyan Chowdhury et al. "Programmable bacteria induce durable tumor regression and systemic antitumor immunity". In: *Nature medicine* 25.7 (2019), pp. 1057–1063.

[14] Candice R Gurbatri et al. "Engineered probiotics for local tumor delivery of checkpoint blockade nanobodies". In: *Science Translational Medicine* 12.530 (2020), eaax0876.

[15] Candice R Gurbatri, Nicholas Arpaia, and Tal Danino. "Engineering bacteria as interactive cancer therapies". In: *Science* 378.6622 (2022), pp. 858–864.

[16] Deborah Nejman et al. "The human tumor microbiome is composed of tumor type–specific intracellular bacteria". In: *Science* 368.6494 (2020), pp. 973–980.

[17] Nuala A O'Leary et al. "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation". In: *Nucleic acids research* 44.D1 (2016), pp. D733–D745.

# 5. Appendix

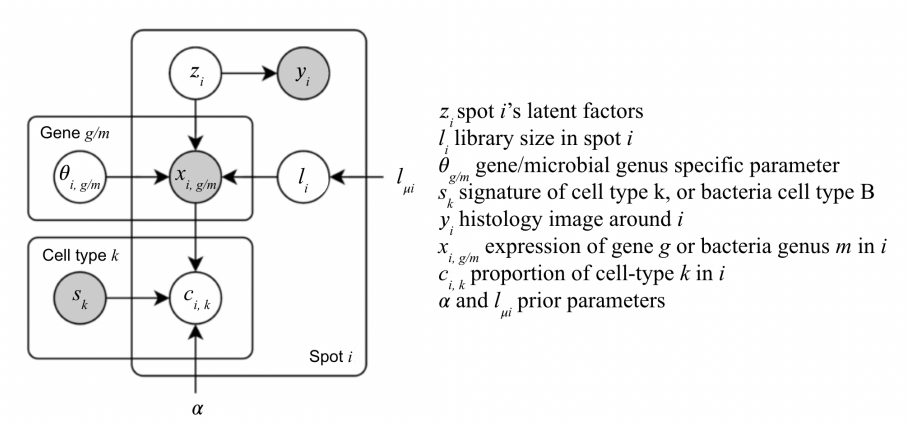


*Figure 5.* Graphical representation of the adapted Starfysh algorithm that incorporates spatial transcriptomics datasets, bacteria taxonomy and histology images. Figure adapted from He et al.

The bacteria-associated archetype was found by calculating the proportion of overlapped spots between the inferred archetypes found through Starfysh and the known cell types using only the cell type signatures. Archetype 12 (Fig. 5) was identified as overlapping the most with the spots containing bacteria and was later used as the bacteria-associated archetype in our analysis.

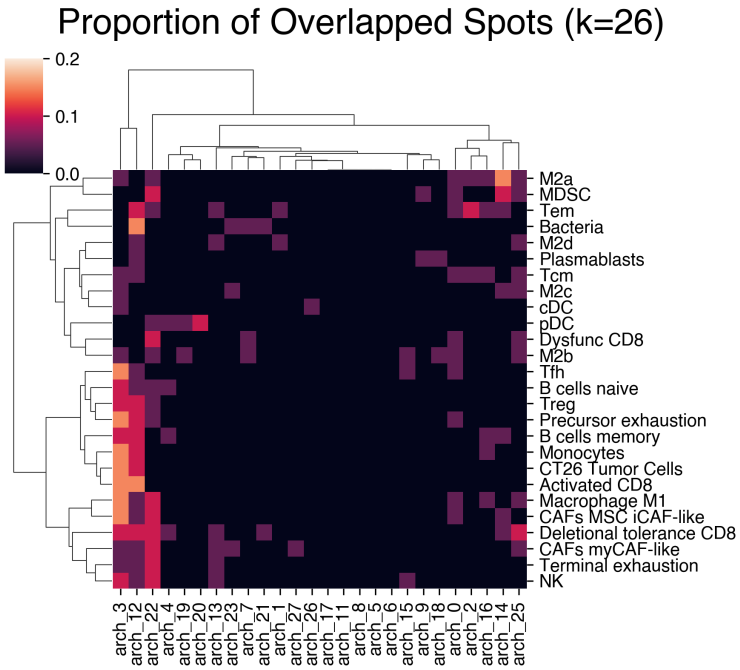


*Figure 6.* Proportion of overlapped spots between the identified archetypes and the cell types given signatures

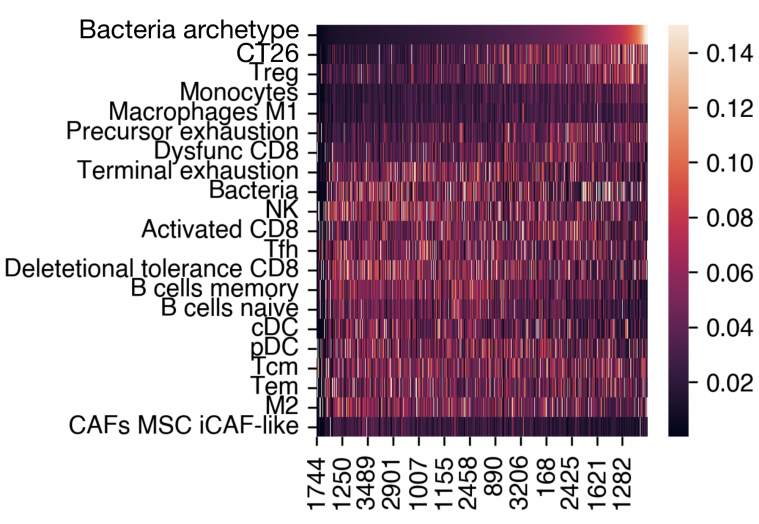


*Figure 7.* Heatmap of inferred cell type proportions in spots ordered by increasing proportion of the bacteria-associated archetype

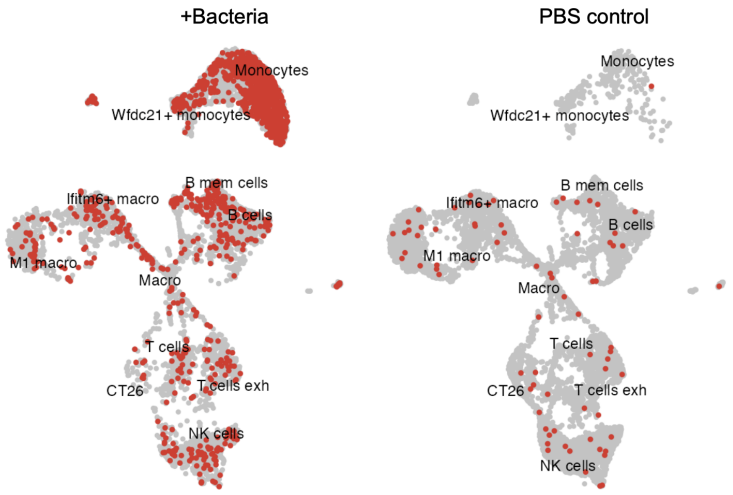


*Figure 8.* UMAP representation of the single-cell datasets highlighting cells with associated bacterial transcripts, left panel is +bacteria condition (intratumoral injection of *E. coli Nislux*) and right panel is PBS control

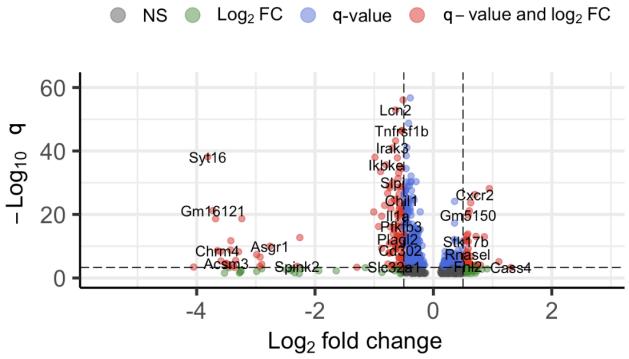


*Figure 9.* Volcano plot of DEGs analyzed as a function of DC1 and +bacteria or -bacteria cells. Upregulated genes are upregulated in bacteria- compared to bacteria+. Q-value (adjusted p-val) is shown on the y-axis.