Denoising sparse microbial signals from single-cell sequencing of mammalian host tissues

Bassel Ghaddar[1*], Martin J. Blaser[2], Subhajyoti De[1]

[1]Center for Systems and Computational Biology, Rutgers Cancer Institute of New Jersey, Rutgers University; 195 Albany St., New Brunswick, New Jersey 08901
[2]Center for Advanced Biotechnology and Medicine, Rutgers University; 679 Hoes Lane West, Piscataway, New Jersey 08854
Correspondence: bcg68@rwjms.rutgers.edu

Abstract
We developed SAHMI, a computational pipeline to identify microbial reads from standard single-cell sequencing of mammalian tissues, offering a framework to study host-microbiome interactions at single cell resolution. SAHMI correctly identifies microbes from different types of infected tissue samples. We also provide guidelines and a resource for detection of likely contaminants and spurious false positive taxonomic assignments.

Main text

The microbiome plays an integral role in healthy development, aging, and multiple disease contexts[1]. However, integrated analyses of host-microbiome ecosystems *in vivo* have been difficult to achieve in part due to the lack of relevant model systems[2] and substantial differences across patients[3,4]. Recently, it has been shown that microbial nucleic acids are captured in polyA-selected RNA sequencing of clinical samples from humans[5–7]. This is not a surprising finding given increasing observations of polyadenylated transcripts in prokaryotes[8,9] and that non-polyadenylated mammalian sequences are routinely captured in RNA-seq[10]. However, in samples from low microbiome-biomass and without matched and co-processed negative controls, denoising true microbial signals and removing contaminating species is a substantial challenge[11–13]. The lack of (1) gold standard contamination controls in most genomic experiments, (2) benchmarking for spurious, false-positive taxonomic assignments derived from human samples, and (3) benchmarking of metagenomic read capture from single-cell RNA sequencing (scRNA-seq) protocols thus often prohibits the use of emerging sequencing data to probe host-microbiome ecosystems.

For these reasons, we developed and benchmarked SAHMI (Single-cell Analysis of Host-Microbiome Interactions), a computational pipeline to identify microbial nucleic acid sequences and remove false positives and contaminants from scRNA-seq of host tissues (**Fig. 1a**). SAHMI denoises and decontaminates the output of taxonomic classifiers such as Kraken2Uniq[14,15]. It identifies true taxa by examining the relationship between the number of total and unique k-mers for each taxon, and it removes contaminants by comparing taxa profiles to an extensive negative

control reference dataset. We demonstrate that SAHMI successfully identifies known infections from scRNA-seq data sets from human tissues and that microbes can be paired and jointly analyzed with somatic cells. SAHMI thus unlocks the potential for retrospective host-microbiome analysis from a wealth of existing scRNA-seq studies.

First, we asked whether known pathogens could be differentially detected in a dose-dependent manner in samples with known infection. We analyzed scRNA-seq data for human samples in which the host had a clinically verified infection[16–18] or in which a pathogen was experimentally introduced[19–22]; these represented a variety of bacteria, fungi, viruses, and tissue types (**Table S1**). We used Kraken2Uniq[14,15] to map all reads to a database of human and microbial genomes. The infection scRNA-seq data contained 10^7-10^{10} reads, of which a mean of 1.3% (median 0.05%, standard deviation 3%) mapped to the microbiome. Across all studies, pathogen reads were successfully identified and were significantly increased in the samples with known or experimental infection with respect to the pathogen load (**Fig. 1b**). To validate that these assignments were not artifacts, we mapped pathogen reads from the clinical samples using STAR[23], a slower but more specific RNA-seq aligner. For all studies, >90% of Kraken2Uniq classified pathogen reads were aligned uniquely by STAR to regions throughout their respective genomes (**Fig. 1c**). These results indicate that true microbial nucleic acid sequences are quantitively captured in scRNA-seq.

Some scRNA-seq barcodes tagging microbial reads also tagged somatic cellular RNA, suggesting that these microbes and cells were co-localized. To determine if barcode sharing reflected the pairing of somatic and microbial cells *in vivo*, we examined the data for known microbe-cell-type specific interactions. Using the same samples from the clinical infection studies, we identified the somatic cell types and highlighted cells that were paired with a pathogen (**Fig. 1d**). Within a cell-type, and across all studies, microbe-associated cells generally clustered together, indicating shared, broad gene expression changes compared to unassociated cells. *Mycobacterium leprae* was most commonly found with T-cells, keratinocytes, and myeloid cells, consistent with its ability to directly infect macrophages and keratinocytes and reflecting the importance of T-cells in the immune responses related to granulomata[24]. *Helicobacter pylori* was mostly associated with a major subset of gastric epithelial cells and with mucosal immune cells. Severe acute respiratory syndrome related coronavirus 2 (SARS-CoV-2) was found broadly in epithelial and immune cells, and especially in alveolar macrophages (**Fig. 1d**). In all studies, pathogen-associated host cells were significantly more likely to be in the same shared nearest-neighbor network with other pathogen-associated cells compared to pathogen-unassociated cells, indicating their transcriptional similarity (Wilcoxon, $p<2e-16$; **Fig. 1e**). These data showcase how the increased precision afforded by single-cell barcoding can localize microbes to specific host cells and enable the examination of cell-type specific gene expression as it relates to the presence of a microorganism.

While these analyses demonstrate the detection of true microbes, a mean of 4735 other species (range 801-7084) were classified in the benchmark datasets, and this number correlated with the total number of reads per study ($r=0.7$, $p=0.08$, **Fig. 2a**). These included common contaminants such as *Mycoplasma* spp. and *Cutibacterium acnes*, but also unexpectedly ubiquitous

species such as *Xanthomonas euvesicatoria* and proteus virus Isfahan. This was especially surprising because five of the eight benchmark studies had pathogens experimentally introduced under strict aseptic technique and suggests that the majority of reported taxa were spurious false-positive assignments. We also observed that reported microbial profiles differed significantly depending on the mapping parameters used. For example, mapping reads from the clinical studies to the microbiome alone without including the host genome led to significantly increased reads that mapped to bacteria in general but to only negligible differences in the number of reads that mapped to the verified pathogen (t-test $p<1e-4$, **Fig. 2b**) – again underscoring the presence of false-positives as well as the importance of including all relevant reference genomes during taxonomic classification. The number of unique k-mer sequences assigned to a taxon has been used to filter false positives; however, we observed a wide range in these values across taxa, and they did not clearly distinguish the known pathogens (**Fig. 2c**).

The issue of false positive taxonomic assignments is well-known[11,25]; these may arise from multiple sources, including sequence homology, sequencing errors, off-target amplification, or mapping errors. To limit such identifications, we posited that in the setting of low-microbiome biomass, a proportional number of total and unique RNA transcripts will be captured for true species. When we examined the pairwise correlations across samples between the numbers of reads, k-mers, and unique k-mers assigned to each species in each benchmark study, we found a wide spectrum of values (**Fig. 2d**), with the true pathogens having extremely high correlation values (median $r>0.92$, $p<2e-16$, **Fig. 2e, Table S1**). The strong correlation between the number of total and unique k-mers also held true for the known pathogens across barcodes, allowing us to identify microbes in individual samples (**Fig. 2f**). These k-mer correlation tests served as a significant filter; while 6063 species were classified in the tested samples, only 1207 species had significant values in all tests.

We next sought to identify contaminants based on the observation that contaminants appear in higher frequencies in negative control samples[12,13]. In the absence of matched controls, we posited that sterile cell line data could serve as a substitute. We profiled the microbiome from publicly available RNA-seq data for 2,491 samples involving >1500 human cell lines representing healthy and diseased tissues from >400 sources from around the world (**Table S2**), thereby creating a negative control resource and identifying common contaminants and false positives (**Table S3**). From these studies, we identified a mean of 1035 species per sample (range: 124-6731). Our k-mer correlation tests again found a range of values for the cell line taxa (**Fig. 2g**). These correlations were significantly weaker than those from the benchmark studies (Wilcoxon, $p<2e-16$) due to the cell line microbiome data being enriched in false positives. The most ubiquitous species included cutaneous microbiota and common environmental or laboratory species (**Table S3**).

Comparing taxa reads counts from the benchmark studies to their distribution in the cell line data using a quantile test clearly distinguished true positive signal and identified background contamination or noise in all studies (Wilcoxon $p<2e-16$, **Fig. 2h**). The known pathogens had significantly higher reads per million microbiome reads compared to what was found in the cell

line data, whereas most taxa did not. Our procedures significantly reduced false positives: only a minority of all reported species (median: 2.8%, range: 0.26-22%) and initially classified microbial reads (median: 12%, range: 3.2-81%) passed both the k-mer correlation and cell line quantile tests (**Fig. 2i**). In the studies with an introduced pathogen (e.g. HSV-1, *S. enterica*), only a median of 3 species per sample passed our pipeline (range: 1-6). To validate that our pipeline enriched for true signal, we used STAR[23] to map reads for a subset of species from the skin leprosy study and found that species that passed both k-mer correlation and cell line quantile tests had significantly more mapped reads than all other initially reported taxa (**Fig. 2j**), despite comprising a minority of reads and species.

These analyses collectively demonstrate how SAHMI can systematically enrich for true taxa and eliminate contaminants and false positives across a range of tissues, enabling retrospective host-microbiome analysis from scRNA-seq data. Furthermore, SAHMI offers a resource to study microbial-cell-type-specific interactions *in vivo*. SAHMI is available on Github (https://github.com/sjdlabgroup/SAHMI).

Methods

SAHMI pipeline for microbiome detection and denoising from scRNA-seq data

We developed SAHMI, a statistical pipeline for detection of microbes and analysis of host-microbiome interactions from scRNA-seq data. After sequencing and taxonomic classifications, SAHMI has two primary functions: (1) it identifies true microbial signal by running correlation analyses across barcodes and samples (k-mer correlation tests), and (2) it filters contaminants and false positives by comparing metagenomic counts to distributions of profiles in negative control samples (cell line quantile test). This enables systematic retrospective identification of microbes in host tissues. Downstream analysis of associations can be done at the sample level or at the level of individual cells for somatic cells and microbes that are tagged with the same cell barcodes.

scRNA-seq and taxonomic classification: metagenomic classification of paired-end reads from single-cell RNA sequencing fastq files can be performed using any k-mer based mapper that identifies a taxonomic ID for each k-mer and read. However, SAHMI is optimized to run with Kraken2Uniq[14,15], which finds exact matches of candidate 35-mer genomic substrings to the lowest common ancestor of genomes in a reference metagenomic database. It is essential that all realistically possible genomes are included as mapping references at this stage, or that host mappable reads are excluded. The required outputs from this step are: a Kraken summary report with sample level metagenomic counts, a Kraken output file with read and k-mer level taxonomic classifications, and raw sequencing fastq files with taxonomic classification for each read, or the equivalent data.

Barcode level signal denoising (barcode k-mer correlation test): SAHMI first extracts microbiome reads from the raw data given their taxonomic IDs and removes reads that contain k-mers that map to the host genome. Next, for each taxon in the sample, SAHMI identifies the corresponding reads and removes reads with less than a default of 50% of the k-mers mapped directly to the taxon or to a parent taxon in its lineage. The scRNA-seq cell barcode is used to

identify reads originating from the same droplet. The number of total and unique k-mers mapping to the taxon or its lineage is then tabulated. For computational efficiency, a default of 1000 barcodes per taxon are randomly sampled. The Spearman correlation between the number of total and unique k-mers across barcodes for each taxon is computed. SAHMI reports the correlation and p-value and recommends removing taxa with non-significant correlations. This enables identification of true microbes in an individual sample.

Sample-level signal denoising (sample k-mer correlation test): the correlation analysis is also conducted across samples when possible. The Kraken report tabulates the total number of reads, minimizers (k-mers), and an estimate of unique k-mer counts for each taxon, or the equivalent data can be obtained from mappers. For each taxon, SAHMI correlates the number of k-mers vs. unique k-mers, reads vs. k-mers, and reads vs. unique k-mers across all samples in a study. True taxa are identified as having significant positive Spearman correlation values and p-values for all three tests.

Identifying contaminants and false positives (cell line quantile test): these can be identified based on the widely observed pattern that contaminants appear at higher frequencies in low concentration or negative control samples[12,13]. We observed that this pattern also extends to false positive assignments. In the absence of experimentally matched negative controls, we provide a negative control resource comprised of microbiome profiles from 2,491 sterile cell experiments from around the world. For each taxon in a test sample, SAHMI compares the fraction of microbiome reads assigned to the taxon [i.e. taxon counts/sum(all bacterial, fungal, viral counts), in reads per million] to the microbiome fraction assigned to the taxon in all cell line experiments. Using the microbiome fraction comparison normalizes for experiments having a varying number of total sequencing reads or varying underlying contamination. SAHMI tests whether the taxon microbial fraction in the test sample is > 99th percentile (by default) of the taxon's microbiome fraction distribution in cell line data using a one-sample quantile test. Taxa whose counts fall within the cell line distribution are identified as below the cell-line noise threshold. Users may choose how stringently to select the quantile threshold for significance testing.

Quantitation of microbes and creating the barcode-metagenome counts matrix: after identifying true taxa, reads assigned to those taxa are extracted and then undergo a series of filters. ShortRead is used to remove low complexity reads (< 20 non-sequentially repeated nucleotides), low quality reads (PHRED score < 20), and PCR duplicates tagged with the same unique molecular identifier and cellular barcode. Non-sparse cellular barcodes can be selected by using an elbow-plot of barcode rank vs. total reads, smoothed with a moving average of 25, and using a cutoff at a change in slope < 10^{-3}, in a manner analogous to how cellular barcodes are typically selected in single-cell sequencing data (CellRanger (10x Genomics), Drop-seq Core Computational Protocol v2.0.0 (McCarroll laboratory)). Lastly, the full taxonomic classification of all resulting reads and the number of reads assigned to each clade are tabulated.

Assembling the negative control cell lines microbiome data

The Sequence Read Archive was queried using the following search: ((((("public"[Access]) AND "rna seq"[Strategy]) AND "transcriptomic"[Source]) AND cell line) AND "Homo sapiens"[orgn:__txid9606] to identify sequencing runs with human cell lines. This resulted in 52,397 sample entries. We then selected for samples with "library selection = cDNA OR library selection = PolyA", and we removed experiments with mouse strain information, experiments involving infection, runs without a submitting center name, and runs with "cell line" designated as "none". From each remaining submitting center, we randomly selected 5 runs. We combined these run IDs with the run IDs for the complete Cancer Cell Line Encyclopedia[26]. We downloaded raw RNA-seq fastq files for these samples and profiled their microbiome using Kraken2Uniq[14,15]. Samples were retained in the data base only if >90% of the reads were classified as human (taxid=9606). We then manually checked each sample's metadata to ensure they did not involve infection or stimulation with microbial agents. The resulting samples and their metadata are tabulated in **Table S2** and cell line microbiome data are available in **Table S3**.

True positive datasets selection, metagenomic mapping, and scRNA-seq data processing

We analyzed scRNA-seq data from patient samples with the following clinically verified infections: *Mycobacterium leprae* (skin)[17], *Helicobacter pylori* (stomach)[18], and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2, bronchoalveolar lavage fluid)[16], in addition to data from scRNA-seq experiments for tissues in which the following pathogens introduced: *Candida albicans* (PBMCs)[20], *Salmonella enterica* (PBMCs)[19], *Mycobacterium tuberculosis* (lung)[21], human alpha herpesvirus 1 (HSV-1) (PBMCs)[27], and human immunodeficiency virus 1 (HIV-1) (PBMCs)[22]. Taxonomic classification was done using Kraken2Uniq[14,15] using a combined reference database including the human genome and all bacterial, fungal, and viral reference genomes recorded in RefSeq as of April 2022. Microbial read alignment was done for a subset of samples using the STAR[23] RNA-seq aligner with the following parameters: alignIntronMax=1 and outFilterScoreMinOverLread=0.05. For the clinical datasets, scRNA-seq data were processed using the standard Seurat[28] pipeline. Cell types were identified by comparing cluster marker genes to the PanglaoDB[29] reference database.

Statistical analyses

All statistical analyses were performed using R version 3.6.1. All p-values were corrected for false-discovery rate (fdr) for multiple hypotheses using the p.adjust function with method= "fdr", unless otherwise stated. The ggpubr package (https://github.com/kassambara/ggpubr) was used to compare group means with nonparametric tests and to perform multiple hypothesis correction for statistics that are noted in the figures. P-values reported as $<2.2\times10^{-16}$ result from reaching the calculation limit for the native R statistical test functions and indicate values below this number, not a range of values. Barcode level analyses (**Fig. 2g, 2j**) were done for the benchmark studies in which reads had identifiable cell barcodes, unique molecular identifiers, and poly-A sequences. Comparisons to the human cell line negative control dataset were done for the benchmark studies that used human cells.

References

1. Cho, I. & Blaser, M. J. The human microbiome: at the interface of health and disease. *Nat. Rev. Genet.* **13**, 260–270 (2012).

2. Glowacki, R. W. P., Engelhart, M. J. & Ahern, P. P. Controlled Complexity: Optimized Systems to Study the Role of the Gut Microbiome in Host Physiology. *Front. Microbiol.* **12**, (2021).

3. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).

4. Ahn, J. & Hayes, R. B. Environmental Influences on the Human Microbiome and Implications for Noncommunicable Disease. *Annu. Rev. Public Health* **42**, 277–292 (2021).

5. Mangul, S. *et al.* ROP: dumpster diving in RNA-sequencing to find the source of 1 trillion reads across diverse adult human tissues. *Genome Biol.* **19**, 36 (2018).

6. Simon, L. M. *et al.* MetaMap: an atlas of metatranscriptomic reads in human disease-related RNA-seq data. *Gigascience* **7**, giy070 (2018).

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

8. Maes, A. *et al.* Landscape of RNA polyadenylation in E. coli. *Nucleic Acids Res.* **45**, 2746–2756 (2017).

9. Hajnsdorf, E. & Kaberdin, V. R. RNA polyadenylation and its consequences in prokaryotes. *Philos. Trans. R. Soc. B Biol. Sci.* **373**, 20180166 (2018).

10. Wang, X., He, Y., Zhang, Q., Ren, X. & Zhang, Z. Direct Comparative Analyses of 10X Genomics Chromium and Smart-seq2. *Genomics, Proteomics Bioinforma.* **19**, 253–266 (2021).

11. Ye, S. H., Siddle, K. J., Park, D. J. & Sabeti, P. C. Benchmarking Metagenomics Tools for Taxonomic Classification. *Cell* **178**, 779–794 (2019).

12. Davis, N. M., Proctor, D. 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).

13. Lusk, R. W. Diverse and Widespread Contamination Evident in the Unmapped Depths of High Throughput Sequencing Data. *PLoS One* **9**, e110808 (2014).

14. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).

15. Breitwieser, F. P., Baker, D. N. & Salzberg, S. L. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. *Genome Biol.* **19**, 198 (2018).

16. Liao, M. *et al.* Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat. Med.* **26**, 842–844 (2020).

17. Ma, F. *et al.* The cellular architecture of the antimicrobial response network in human leprosy granulomas. *Nat. Immunol.* **22**, 839–850 (2021).

18. Zhang, P. *et al.* Dissecting the Single-Cell Transcriptome Network Underlying Gastric Premalignant Lesions and Early Gastric Cancer. *Cell Rep.* **27**, 1934-1947.e5 (2019).

19. Saliba, A. E. *et al.* Single-cell RNA-seq ties macrophage polarization to growth rate of intracellular Salmonella. *Nat. Microbiol.* **2**, 1–8 (2016).

20. Muñoz, J. F. *et al.* Coordinated host-pathogen transcriptional dynamics revealed using sorted subpopulations and single macrophages infected with Candida albicans. *Nat. Commun.* **10**, (2019).

21. Pisu, D. *et al.* Single cell analysis of M. tuberculosis phenotype and macrophage lineages in the infected lung. *J. Exp. Med.* **218**, e20210615 (2021).

22. Golumbeanu, M. *et al.* Single-Cell RNA-Seq Reveals Transcriptional Heterogeneity in Latent and Reactivated HIV-Infected Cells. *Cell Rep.* **23**, 942–950 (2018).

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

24. Mi, Z., Liu, H. & Zhang, F. Advances in the Immunology and Genetics of Leprosy. *Front. Immunol.* **11**, (2020).

25. McIntyre, A. B. R. *et al.* Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biol.* **18**, 182 (2017).

26. Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).

27. Wyler, E. *et al.* Single-cell RNA-sequencing of herpes simplex virus 1-infected cells connects NRF2 activation to an antiviral program. *Nat. Commun.* **10**, 4878 (2019).

28. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

29. Franzén, O., Gan, L.-M. & Björkegren, J. L. M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database* **2019**, baz046 (2019).

Supplementary Materials
Table S1. Summary of benchmark studies analyzed.
Table S2. Summary of cell line experiments analyzed.
Table S3. Cell lines microbiome reference

Acknowledgments
We acknowledge the Office of Advanced Research Computing (OARC) at Rutgers, The State University of New Jersey for providing access to the Amarel cluster URL: https://it.rutgers.edu/oarc.

Funding
National Institutes of Health grant R21CA248122 (SD)
National Institutes of Health grant U01 AI22285 (MJB)
Sergei Zlinkoff Foundation (MJB)
Canadian Institute for Advanced Research (MJB)
National Institutes of Health, National Center for Advancing Translational Sciences, Rutgers Clinical and Translational Science Award TL1TR003019 (BG)

Author contributions
BG and SD conceived the study. BG designed and performed all data analyses. BG, MJB, SD interpreted the data and wrote and revised the manuscript.

Competing interests
MJB declares that he serves on the Scientific Advisory Board of Micronoma, Inc. BG and SD have jointly filed PCT patent applications PCT/US2022/025829 and PCT/US2022/025832.

Data and materials availability:
The cell lines microbiome negative control dataset is available in **Supplementary Table S3**. The SAHMI pipeline is available on our Github: https://github.com/sjdlabgroup/SAHMI.

Figure Legends
Figure 1. Host scRNA-seq captures true microbial reads. a, Schematic of SAHMI pipeline. SAHMI identifies true positive taxa using a k-mer correlation test and identifies false positives and contaminants by comparing taxa distributions to a negative control reference. **b,** Differential detection of known pathogens in infected samples in multiple studies. Boxplots show median (line), 25th and 75th percentiles (box) and 1.5xIQR (whiskers). Points represent outliers; all Wilcoxon testing. rpm, reads per million classified microbiome reads. **c,** Histograms of genomic mapping positions for the known pathogens in the clinical studies. Reads map throughout the genomes. **d,** Uniform manifold approximation and projection (UMAP) plots of somatic cells from the clinical studies identifying cell types and cell barcodes paired with the known pathogen (red color).

Microbe-associated cells cluster together. Tissue sources: left, gastric biopsy; middle, skin; right: bronchoalveolar lavage fluid. **e,** Boxplots of the probability a pathogen-associated or unassociated host cell is in the same shared nearest-neighbor network as a pathogen-associated cell. Boxplots are as in (b).

Figure 2. Distinguishing true signal from false positives and contaminants. a, Scatter plot showing the total number of sequencing reads and species detected in each study. Blue, experimentally introduced pathogen; red, existing infection in a human tissue. **b,** Boxplots showing significantly increased reads assigned to bacteria when the human genome is not included as a reference during taxonomic classification. Boxplots show median (line), 25th and 75th percentiles (box) and 1.5xIQR (whiskers). Points represent outliers; t-tests; ****, p<1e-3; ***, p<1e-2; ns, not significant. **c,** Histogram of the number of unique k-mers per sample assigned to the known pathogens and to all detected species in the benchmark studies. **d,** Scatter plot of the k-mer correlation tests for species in the benchmark studies. Each point represents a species. Correlations are run across samples in a study. x-axis, Spearman correlation value between #k-mers vs. #unique k-mers; y-axis, correlation value between #k-mers vs. #reads; color, correlation value between #reads vs. #unique k-mers. Lines represent contour density. **e,** Scatter plot of the number of total and unique k-mers for the known pathogens. Each point represents a sample. The points are colored by study as detailed below the plot. **f,** Scatter plot of the number of total and unique k-mers for the known pathogens. Each point represents an individual barcode. Points are colored as in (e). **g,** Similar to (**d**) but for taxa detected in cell line experiments. **h,** Overlaid histograms of reads per million microbiome reads assigned to the known pathogens (top row) and to example contaminants (bottom row) in the benchmark studies as well as in the cell line data. Bars are colored by study of origin. **i,** Bar plots of relative proportions of species (left panel) and reads (right panel) that passed or failed the k-mer correlation and cell-line quantile tests. **j,** Boxplots showing the number of reads per taxa mapped by STAR for microbial reads from the leprosy study grouped by k-mer correlation and cell line quantile tests results. Colors are as in (i) and boxplots are as in (b). Wilcoxon testing.

Figure 1

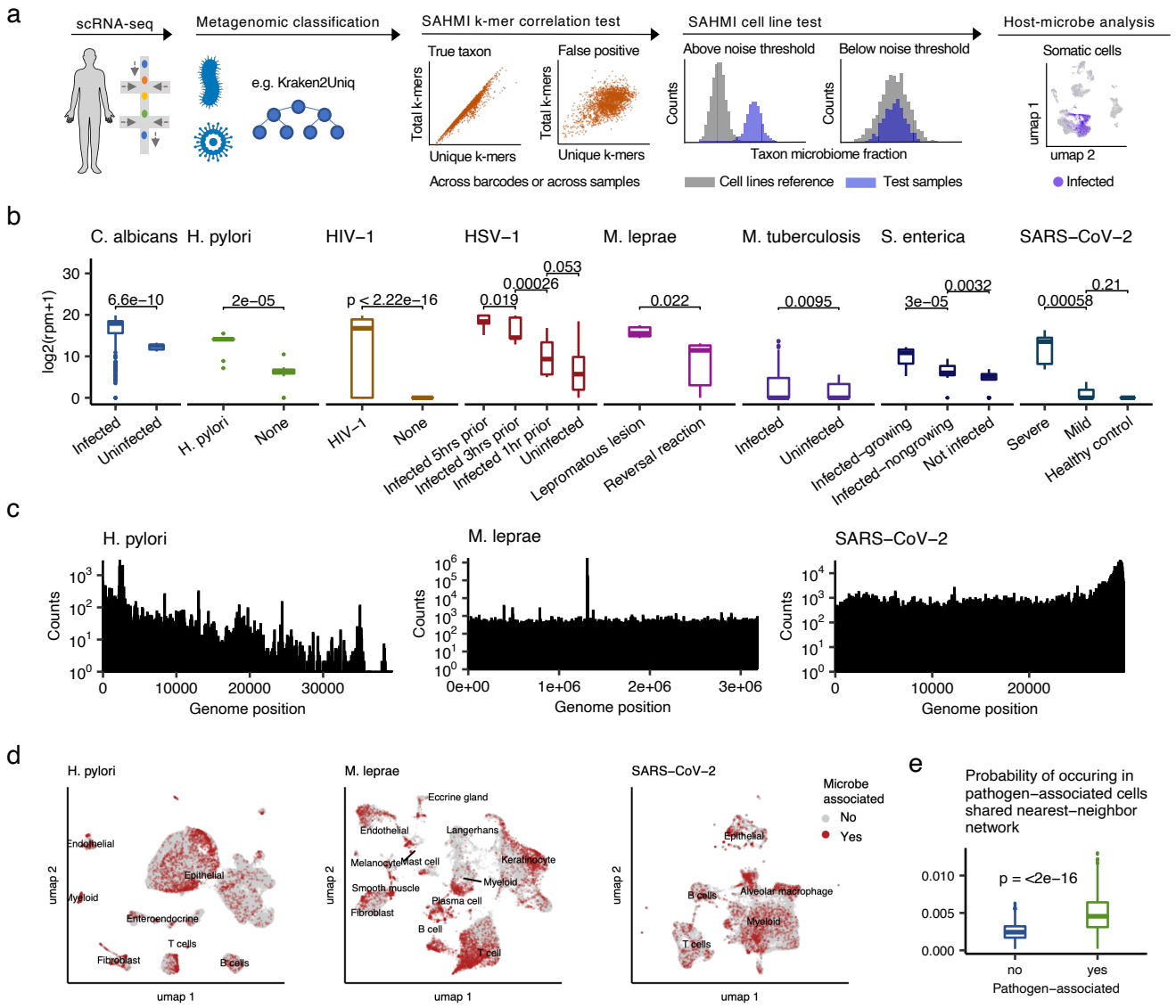
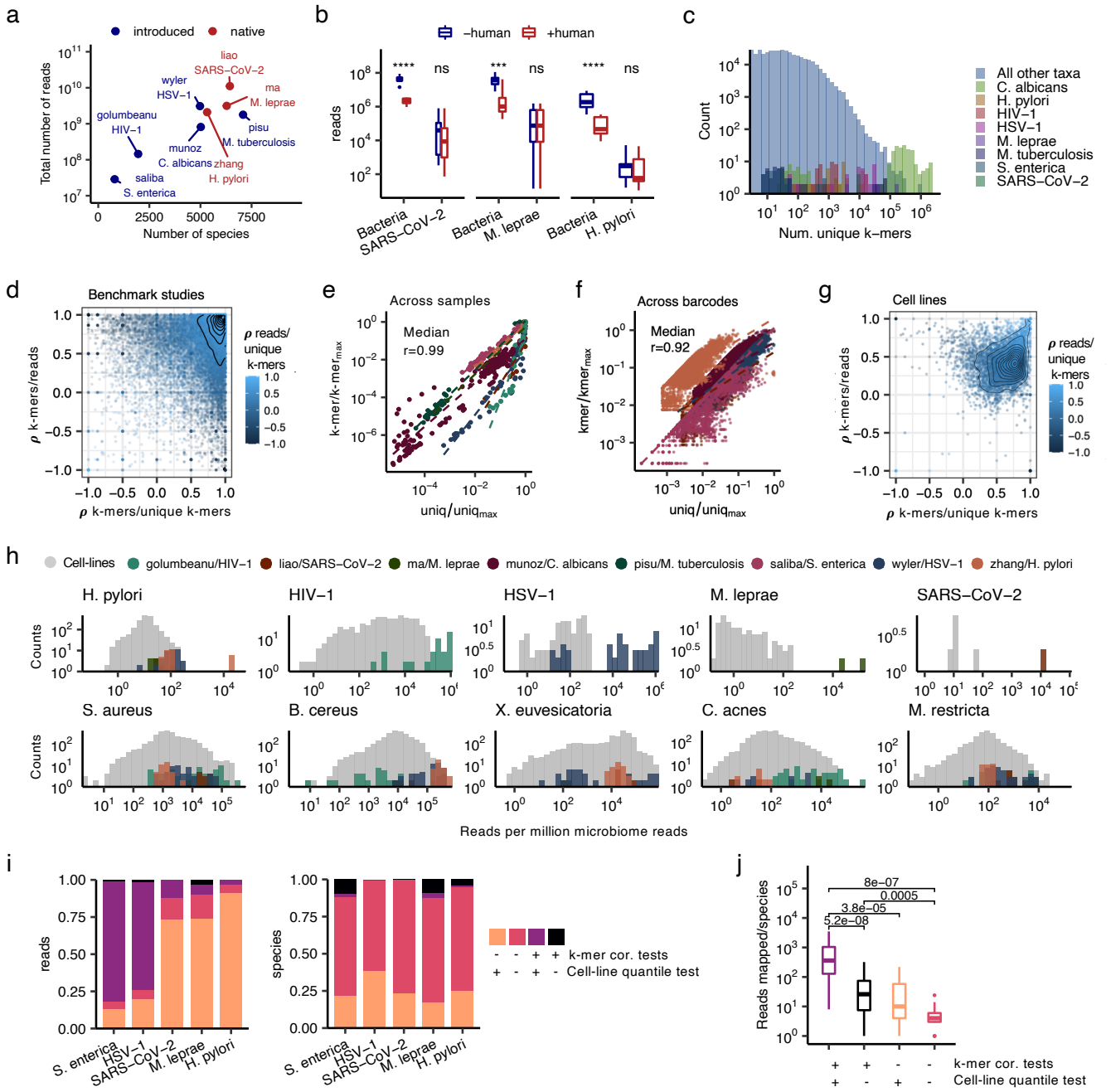

a
scRNA-seq
Metagenomic classification
e.g. Kraken2Uniq
SAHMI k-mer correlation test
True taxon
False positive
Total k-mers
Unique k-mers
Across barcodes or across samples
SAHMI cell line test
Above noise threshold
Below noise threshold
Counts
Taxon microbiome fraction
Cell lines reference
Test samples
Host-microbe analysis
Somatic cells
umap 1
umap 2
Infected
b
C. albicans
H. pylori
HIV−1
HSV−1
M. leprae
M. tuberculosis
S. enterica
SARS−CoV−2
log2(rpm+1)
6.6e−10
2e−05
p < 2.22e−16
0.019
0.00026
0.053
0.022
0.0095
3e−05
0.0032
0.00058
0.21
Infected
Uninfected
H. pylori
None
HIV−1
Infected 5hrs prior
Infected 3hrs prior
Infected 1hr prior
Lepromatous lesion
Reversal reaction
Infected−growing
Infected−nongrowing
Not infected
Severe
Mild
Healthy control
c
H. pylori
M. leprae
SARS−CoV−2
Counts
Genome position
d
H. pylori
M. leprae
SARS−CoV−2
umap 1
umap 2
Endothelial
Epithelial
Myeloid
Enteroendocrine
T cells
Fibroblast
B cells
Eccrine gland
Langerhans
Melanocyte
Mast cell
Keratinocyte
Smooth muscle
Plasma cell
B cell
T cell
Alveolar macrophage
Microbe associated
No
Yes
e
Probability of occuring in pathogen−associated cells shared nearest−neighbor network
p = <2e−16
no
yes
Pathogen−associated

Figure 2


a
introduced
native
Total number of reads
Number of species
liao
SARS−CoV−2
wyler
HSV−1
ma
M. leprae
golumbeanu
HIV−1
pisu
M. tuberculosis
munoz
C. albicans
zhang
H. pylori
saliba
S. enterica
b
−human
+human
reads
****
ns
***
ns
****
ns
Bacteria
SARS−CoV−2
Bacteria
M. leprae
Bacteria
H. pylori
c
Count
Num. unique k−mers
All other taxa
C. albicans
H. pylori
HIV−1
HSV−1
M. leprae
M. tuberculosis
S. enterica
SARS−CoV−2
d
Benchmark studies
ρ k-mers/reads
ρ k-mers/unique k-mers
ρ reads/ unique k-mers
e
Across samples
Median r=0.99
k-mer/k-mer_max
uniq/uniq_max
f
Across barcodes
Median r=0.92
kmer/kmer_max
uniq/uniq_max
g
Cell lines
ρ k-mers/reads
ρ k-mers/unique k-mers
ρ reads/ unique k-mers
h
Cell-lines
golumbeanu/HIV−1
liao/SARS−CoV−2
ma/M. leprae
munoz/C. albicans
pisu/M. tuberculosis
saliba/S. enterica
wyler/HSV−1
zhang/H. pylori
H. pylori
HIV−1
HSV−1
M. leprae
SARS−CoV−2
S. aureus
B. cereus
X. euvesicatoria
C. acnes
M. restricta
Counts
Reads per million microbiome reads
i
reads
species
S. enterica
HSV−1
SARS−CoV−2
M. leprae
H. pylori
k-mer cor. tests
Cell-line quantile test
j
Reads mapped/species
8e−07
0.0005
3.8e−05
5.2e−08
k-mer cor. tests
Cell-line quantile test