Single Cell Transcriptomics Reveals the Hidden Microbiomes of Human Tissues

Gita Mahmoudabadi[1], Tabula Sapiens Consortium[2], Stephen R. Quake[1,3,4,*]

[1]Department of Bioengineering, Stanford University, Stanford CA, USA
[2]Authors and their affiliations are listed at the end
[3]Department of Applied Physics, Stanford University, Stanford CA, USA
[4]Chan-Zuckerberg Biohub, Stanford CA, USA
*Corresponding author. Email: steve@quake-lab.org

Abstract

The human microbiome has been studied extensively across sites in the body that are readily accessible to sampling. Internal organs and tissues, however, have remained largely unexplored and, in the absence of infectious disease, are widely assumed to be free of microorganisms. Using single-cell transcriptomic data from the Tabula Sapiens spanning 15 human organ donors, 20 tissues, 400,000+ annotated cells, 100+ cell types, and ~70 billion sequences, we created an atlas of the healthy human tissue microbiome with cell type resolution. In order to construct this atlas, we developed a computational pipeline which identifies high-confidence microbial sequences from three domains of life within the sea of human transcripts in individual cells, and show that this is a non-trivial search for one in a million sequences. Together with data from a separate validation cohort of 8 additional donors, we identify more than a thousand diverse bacterial, viral and fungal species in human tissues. In characterizing the human tissue microbiome, we demonstrate that individuals harbor unique species associated not just with each tissue but often with particular cell types, and that many of these species remain uncharacterized. Moreover, combining our data with the Human Microbiome Project, the tumor microbiome dataset by Nejman *et al*., and the PATRIC database, we map the likely microbial flow routes from external-facing microbiomes to internal tissues and tumors, and show the existence of many unexpected routes taken by both commensals and pathogens. We find that ~30% of bacterial species found in tumors are detectable across healthy tissues, suggesting that tumor microbiomes are in part sourced from healthy tissues, even those from distant sites. Increasing the resolution of sampling from tissues to cell types, we quantify the microbial load and diversity across different cell types to reveal a network of host cell type and microbe associations. For example, we identified traces of both latent and active Epstein Barr Virus infections in various cell types such as splenic plasma cells. Broad exploration of the healthy tissue microbiome may provide insights which ultimately are of clinical importance.

Introduction

Microbial life contains a vast repertoire of biochemical pathways that enable it to colonize even the most extreme environments on the planet. It is no wonder that the human body, an energy-rich environment for trillions of our own cells, would be host to an even larger number of highly diverse microbes. Historically, understanding of microbes in connection to human health has emerged from waves of encounters with devastating pathogens such as *Mycobacterium tuberculosis*, *Vibrio cholerae*, and most recently, SARS-CoV2. In the last two decades our view of human-microbe interactions has been transformed through

increasingly powerful sequencing technologies which have enabled broad surveys of the microbial world[1,2,3,4]. As such, previously held dogma have come unraveled, for example the assertion that internal organs such as the healthy human lung are sterile environments[5].

The study of the human microbiome has seen an explosion of data connecting communities of microbes or individual species to a variety of metabolic[6,7], neurological[8,9], cardiovascular[10,11], and autoimmune[12,13] diseases, among others. The vast majority of microbiome studies, however, have delineated microbial communities at sites easily accessible to sampling such as the human intestinal tract via feces, the mouth, and the skin. Less accessible internal sites, on the other hand, are largely assumed to be free of microorganisms, as they are protected by layers of epithelial and endothelial tissue. However, a few recent studies have challenged this assumption. For example, human tumors were shown to contain tumor-specific microbial sequences, providing a possible diagnostic approach for cancer detection[14]. Another study similarly identified the presence of tumor-specific microbiomes including intracellular bacteria[15].

To fully understand the significance and role of tumor microbiomes[16,17,18] and microbiomes across various disease states requires that we also study healthy human tissue microbiomes. To address this, we assembled the first atlas of human tissue and cell type microbiomes by mining one of the largest single-cell transcriptomic datasets of human tissues, the Tabula Sapiens[19] (TS), encompassing 400,000+ annotated human cells, spanning 100+ cell types and 20 tissues, originating from 15 human organ donors (**Figure 1**). In absence of any available pipelines for the identification of microbial sequences in single-cell transcriptomic datasets, we created a computational pipeline called SIMBA or **S**ingle-cell **I**terative **M**ulti-biome **B**LAST **A**nalysis, which identifies high-confidence viral, bacterial and fungal sequences in the vast sea of human transcripts within individual cells (**SI Figure 1**). This multi-biome search was inspired by the fact that many microbiome studies to date have focused solely on characterizing bacteria, often overlooking the important contributions of viruses, fungi and other groups of microbes[20,21,22].

We explored the intersection of species in healthy tissues and those found at tumor sites and at sites accessible to sampling, or what we will refer to as “external-facing microbiomes” (e.g. the gastrointestinal, oral and skin microbiomes), to chart out possible microbial flow routes throughout the body. We report the inter-and intra-donor variability of tissue microbiomes, and characterize microbial load and diversity across cell types. Moreover, using a cohort of 8 additional donors, we devised an assay to extract virus-like and bacterial-like particles from human tissues. Using DNA sequencing, fluorescence, and electron microscopy of these extracted particles, we validate our findings from the

transcriptomic dataset, and demonstrate that the human tissue and cellular microbiomes, beyond cell-free or capsid-free nucleic acids, also includes intact organisms.

Results

SIMBA, a computational pipeline for the identification of bacteria, viruses and fungi in single cells, which reduces the computational cost of high-confidence alignments by 100-fold

Up to now it has been challenging to identify microbial sequences in single cell transcriptomic datasets. A straight-forward approach would be to first eliminate all sequences mapping to the human genome and to align remaining sequences against one of the largest sequence databases, such as NCBI's nt database, in order to identify the most likely organism of origin for each sequence. However, this approach is computationally infeasible given the sheer sizes of both the TS dataset, even after the removal of human transcripts (~6 Billion sequences), as well as the nt database (~70 Million sequences at the time of download). Thus, using BLASTn[23], which is considered the gold standard of alignment tools[24], would require ~3 x 10^8 CPU hours, equivalent to ~$9 million using Amazon's compute instances.

To lower this computational burden, we created a lean human-microbiome centric database for bacteria and fungi which we call microbDB (see Methods) and used as a first-pass filter to select sequences with the highest likelihood of being microbial in origin (**SI Figure 1**). Additionally, through subsampling, we created leaner versions of the publicly available bacterial and fungal RefSeq databases to use as second-pass filters. Finally, surviving sequences that produced significant alignments or "hits" (E-value < 10^{-5}) against these databases were aligned against the nt dataset. Approximately 19 million sequences produced significant hits against nt. Importantly, half of these sequences mapped back to the human genome, and only 12% mapped to the biomes of interest: viral, bacterial and fungal (**SI Figure 1**). This relatively low percentage highlights the critical importance of our approach in having multiple checkpoints for the removal of human and other non-microbial reads. It confirms that the use of custom databases independent of validation against more comprehensive databases such as nt - a strategy often used in microbial metagenomic studies to lower computational cost - would indeed produce a high percentage of false positives, at least in the context of human tissues where most sequences are human in origin. In summary, SIMBA's iterative alignment strategy reduced the computational cost of high-confidence microbial alignments by 100-fold.

In addition to existing quality control steps implemented through SIMBA, we imposed additional stringent filters based on alignment length and percent identity to select only the highest-confidence alignments. The purpose of these more stringent filters applied at this stage as opposed to during alignments to custom databases, was to ensure a more permissive search that would capture, by way of similarity or homology, more species than were included in the custom databases. For example, we were able to obtain hits against an order of magnitude more unique bacterial species than were included in the custom database used for bacteria. Furthermore, we designed a bioinformatic decontamination pipeline for single cell experiments (see Methods). Considering that the starting number of sequences in the TS dataset was ~70 billion and the final number of sequences passing through all filters was ~30,000, we conclude that finding high-confidence microbial sequences in human cellular transcriptomes is a non-trivial search for one in a million sequences.

The human tissue microbiome atlas reveals numerous species of microbes spanning three domains of life, including potentially novel species

Binning sequences by donor and tissue type, we summarize their breakdown into high-level taxonomic categories **(Figure 1, SI Figure 2, SI Table 1)**. Numerous species spanning three domains of life are present across 15 donors and 20 tissue types, with bacteria having the highest prevalence of the three domains. Bacterial orders form substantially different patterns of occurrence across different donors and tissues. Unlike bacterial sequences, which map largely to commensal species, viral sequences mapped mostly to the Epstein Barr Virus (EBV) - a ubiquitous and persistent human pathogen with very limited transcriptional footprint during its latency compared to its reactivation[25,26]. As we will discuss in more detail, this dichotomy in the transcriptional load of EBV during its latency and reactivation likely gives rise to detectable transcripts in some donors while appearing to be missing from others. Finally, fungal sequences are generally present with lower relative abundances compared to bacteria. This is consistent with previous estimates of fungi at external-facing microbiomes, depicting them significantly outnumbered by bacteria[21,27,28].

Bacterial sequences are derived from a range of phyla, most notably Actinobacteria, Firmicutes, and Proteobacteria. Among these sequences, ~20% are ribosomal RNA (rRNA), ~30% of which belong to uncultured or uncharacterized bacteria, representing potentially novel species and even phyla **(SI Figure 3, SI Table 2)**. This relatively large percentage of ribosomal sequences is expected as rRNA is the most abundant form of RNA in bacterial cells, and its abundance is tightly coupled to the bacterial growth rate[29,30], accounting for up to 85% of the RNA pool in rapidly growing bacterial cells in culture[31,32]. Furthermore, our finding of bacterial rRNA across numerous human tissue types is in

agreement with a recent study of several tumor types, wherein 16S rRNA was detected using both sequencing and imaging[15].

Mapping the possible microbial flow routes from external-facing microbiomes to internal tissues and tumors

To explore the human tissue microbiome at a more granular level, we show a Sankey plot (**Figure 2.a**), wherein each one of the ~30,000 surviving microbial sequences is represented by a line, the path of which can be traced to obtain the tissue type of origin, taxonomic categorization, as well as additional information gathered from three external databases. The Human Microbiome Project (HMP) database[33] was used to identify species that had previously been observed at external-facing microbiomes from the gastrointestinal tract, oral cavity, ear, nose, airways, urogenital tract, and the skin. Approximately 20% of bacterial species and 80% of orders that comprise external-facing microbiomes are also detected in tissues (**SI Figure 4**).

With the assumption that microbial flow occurs from external to internal sites, we depict possible flow routes for numerous bacterial species (**Figure 2.a**) and show that various external and internal sites are connected through their shared species (**SI Figure 4, SI Figure 5**). Surprisingly, not all species found at internal sites originate from the gastrointestinal tract. Other external-facing microbiomes appear to contribute many species to internal organs and tissues (**SI Figure 4**). To illustrate a few microbial flow paths, we detected *Treponema denticola,* an oral bacterium commonly associated with periodontal disease, in the heart. This observation is in line with studies that have linked *Treponema denticola* to atherosclerosis through its invasion of aortic tissue[34–36] (**SI Figure 5**). *Streptococcus pyogenes*, another oral bacterium whose infection can lead to rheumatic heart disease[37], was also detected in the heart (**SI Figure 5**). While various members of the oral microbiome have been associated with cardiovascular diseases[38,39], we find that the heart is certainly not the only destination for species normally found in the oral cavity. For example, we found *Streptococcus oralis, Rothia dentocariosa,* and *Porphyromonas gingivalis* not only in the salivary gland tissue as might be expected, but also in tissues from the thymus, liver, pancreas, trachea, bone marrow and others (**SI Figure 5**). Sequences mapping to inhabitants of other external-facing microbiomes were similarly found in various tissues (**Figure 2.a**, **SI Figure 5**). It remains to be discovered to what extent species from external-facing microbiomes can colonize these distant sites.

We further explored the bacterial species shared between healthy tissues (TS dataset) and tumors (Nejman *et al.*[15]) of the pancreas, breast, colon, lung, skin, ovaries, bone, and the brain (**Figure 2.a, SI Figures 6-8**). Approximately 30% of bacterial species and ~70% of bacterial orders found previously in tumors also appear in healthy human tissues (**SI**

Figure 6, SI Figure 7, SI Figure 8). We created an abundance heatmap of these species across healthy tissues (**SI Figure 6.b, SI Table 3**). Several species that were previously identified as highly enriched in tumors including *Klebsiella pneumoniae* and *Fusobacterium nucleatum* are also present across various healthy tissues such as those from the pancreas, the uterus, and the bone marrow (**SI Figure 6.c**). Bacterial orders missing from healthy tissues but appearing in tumors include *Gemellales*, *Streptophyta*, and *Selenomonadales*, and several others (**SI Figure 7.a**). The relatively large overlap between healthy tissue microbiomes and tumor microbiomes suggests that tumor microbiomes are sourced from healthy tissue microbiomes, even possibly from tissues at distant sites from the tumor (**SI Figure 7.b**).

We mapped the possible microbial flow routes from external-facing microbiomes to healthy tissues and tumors by focusing on species at the intersection of the TS, HMP and the Nejman *et al.* datasets (**SI Figure 9, SI Figure 10, SI Table 4**). Many interesting flow routes appear to exist with varying degrees of complexity. As an example of a simple flow route, *Aggregatibacter segnis*, appears to originate from the oral microbiome (HMP). It is detected in healthy salivary gland tissue (TS) and is found in breast tumors (Nejman *et al.*). *Enterobacter cloacae*, on the other hand, an inhabitant of the GI tract (HMP) is not only found in colon tumors but also in pancreatic, breast, and GBM tumors. Its pervasive signature in tumors makes sense in light of its pervasive presence across healthy tissues in the TS dataset as it is found in the pancreas, blood, bone marrow, and lymph nodes.

Additionally, we used the PATRIC database[40] to get further information about the pathogenicity potential of bacterial species (**Figure 2.a, SI Figure 10, SI Figure 11**). Interestingly, we found ~30% of species shared between all three datasets to be potentially pathogenic, compared to only 4% of species within the entire TS dataset. We provide a list of these pathogens with presence in both healthy tissues and tumors (**SI Figure 11**). Using the PATRIC database, we provide the antibacterial resistance profiles against 60 different antibiotics, including those for three of the potential pathogens found in both tumors and healthy tissues, namely *Staphylococcus aureus*, *Klebsiella pneumoniae*, and *Pseudomonas aeruginosa* (**Figure 2.a, SI Figure 12, SI Table 5**).

Human tissue microbiomes are tissue- and individual-specific

To explore the inter-and-intra donor tissue microbiomes at finer taxonomic resolution, we computed the shared species within and across individuals. Whether considering the tissues of one donor or the collective tissue microbiomes of many donors, tissue microbiomes appear to have many overlapping species **(Figure 2.b, SI Figure 13)**. However, when accounting for all species, species unique to each sample significantly outnumbered those that are shared (**Figure 2.c, SI Figure 14**). Furthermore, we compared species-level microbiomes in a pairwise fashion using the Spearman

correlation (**SI Figure 15**), revealing tissue microbiomes across and within each individual to be uncorrelated, and thus both individual- and tissue-specific. The personal nature of human tissue microbiomes most likely reflects back the highly individual-specific microbiomes that exist at external-facing microbiomes[41,1,42,43,22]. As we also demonstrate through the overlap between species from different isolation sites in the HMP database **(SI Figure 16)**, external-facing microbiomes have limited overlap at the species level. Moreover, as we reduce the taxonomic resolution from species to orders and phyla, the correlations between and within individuals increase significantly (**SI Figure 15**), suggesting that high-level tissue microbiome compositions are largely conserved.

Increasing the resolution of tissue microbiomes to delineate cell type microbiomes

Many of the microbial sequences can be traced back to cell types that have been annotated by the Tabula Sapiens Consortium. Merging cell type annotation data with cellular barcodes provides an unprecedented view of each tissue's microbiome at the level of cell types and even individual cells. We wondered how different cell types compare in terms of their microbial load. Thus, we quantified single-cell microbial load as $\log_2(h_c/g_c)$, where h_c represents the number of microbial sequences detected in cell *c* and is normalized by the number of host genes detected in that cell, g_c, to account for cell-to-cell variation in sequencing coverage.

Ranked distributions of single-cell microbial loads (**Figure 3.a**), reveal that cell types with the largest microbial load include plasma B cells, pancreatic acinar cells, enterocytes, and macrophages, among other cell types, especially those from the immune compartment. The over-representation of immune cell types is expected as these cells, with the ability to survey tissues, are the first to encounter microbial organisms. Enterocytes are also the first line of barrier against the trillions of microbes housed in the human intestinal tract. At first glance, the finding of microbial sequences in pancreatic acinar cells and in the pancreas more generally, may be the most surprising observation, however the pancreas is an organ with a physical connection to the small intestine. This connection has been previously shown to enable the passage of an intestinal fungal species to the pancreas where it promotes pancreatic adenocarcinoma[44], and pancreatic adenocarcinoma tumors have been shown to harbor microbiomes that are predictive of long-term survival[17].

Network of human cell types in association with microbes reveals cell types with greatest microbial diversity

To explore the diversity of cellular microbiomes, we built a bipartite network of cell types stratified by tissue of origin (source nodes) and their associated microbial species (target

nodes) (**Figure 3.b**). Cell types in each tissue are weighted by the number of unique microbial species, normalized by the number of cells sequenced. Hence, in this network of host-microbe associations, hubs represent cell types with the greatest microbial diversity. Examples of these hubs include bone marrow macrophages, bladder B cells, blood monocytes and neutrophils, muscle endothelial cells, as well as acinar cells from the pancreas and the salivary glands. As with microbial load measurements, immune cell types are over-represented in their diversity of microbial associations. Salivary gland acinar cells, similar to pancreatic acinar cells, are cell types that are in close proximity to a large reservoir of microbes and therefore expected to experience diverse associations. Cellular microbiome network depictions such as this open a window into host cell and microbe interactions in their native context which would not be captured in bulk tissue sequencing experiments. An example of an important association is between EBV and splenic plasma cells.

Detecting the transcriptional footprints of a persistent viral pathogen and an ancient viral relic in human cells

EBV was the first oncovirus to be discovered and has since been shown to cause Burkitt's, Hodgkin's, NK and T cell lymphomas as well as nasopharyngeal and gastric carcinomas, thus accounting for ~200,000 (or 2% of) cancer deaths annually worldwide[45]. In absence of an approved EBV vaccine, currently most primary infections occur in children and adolescents, such that by adulthood a significant fraction of the population is latently infected[46,47,48]. Additionally, a recent longitudinal study of 10 million young adults has shown a ~30 fold increase in the risk of multiple sclerosis after EBV infection, placing EBV as one of the leading potential causes of multiple sclerosis[49].

During the latent phase of its lifecycle, EBV suppresses the transcription of most of its genome[26] and replicates infrequently and in synchrony with the host cell while utilizing the host polymerase[50]. In our cohort, we detected EBV sequences in three seropositive donors, and across various tissues including the lung and the bladder, with a substantial number of sequences emerging from lymphoid organs such as the spleen, where many of the susceptible host cell types reside (**Figure 4.a**). Consistent with the known aspects of EBV tropism[51], we found that most sequences originate from cell types belonging to immune and epithelial compartments.

Across distributions of EBV single-cell load in different cell types (measured for donor 2 where the majority of EBV sequences originate from), plasma cells outrank other cell types (**Figure 4.b**). Applying BLASTx to identify protein coding sequences, we primarily found those expressed during the lytic phase (**Figure 4.c**): membrane glycoprotein (BILF1), DNA synthesis and genome packaging protein (BALF3)[52], and the viral DNA

polymerase (BALF5), which is used to replicate the EBV genome exclusively during the lytic phase[50]. Notably, we observe BZLF1 expression, an important transcriptional switch that initiates the expression of various lytic genes[53]. Although EBV sequences are present across a variety of cell types, only plasma cells, B cells and NK cells are positive for lytic transcripts, consistent with the fact that these cell types are known susceptible host cell types for EBV. Considering that the latent phase of EBV life cycle is characterized by the expression of non-coding RNAs, many of which suppress lytic genes[45], absence of coding lytic transcripts suggests that other cell types are experiencing latent infections. Alternatively, they may have uptaken transcripts passively by virtue of physical proximity to sites of lytic infections, or actively through their involvement in the clearance of lytically infected cells.

In addition to EBV, we detect the presence of human endogenous retroviruses (HERVs). Introduced numerous times during the evolutionary history of primates, HERV genomes, which are mostly defunct, occupy 8% of the human genome[54]. The HERV K (HML-2) family, representing the latest insertions into the human genome, is the only family of HERVs that is transcriptionally active. Differential expression of HERV K genes has been detected across several cancers and is thought to contribute to tumor progression[54,55,56]. Additionally, HERV K expression and even virus-like particles have been detected during human blastocyst formation[57]. Contrary to the claim that HERV K expression is silenced in all but a few special contexts[58], we identified HERV K (HML-2) transcripts across five donors, spanning a diverse range of healthy tissues (e.g. skin, fat, spleen and the salivary glands) and cell types (**SI Figure 17.a**). These sequences map to the three canonical retroviral genes in the HERV K genome, namely pol, gag and env (**SI Figure 17.b**). HERV K expression in human donors beyond tissues and cell types previously investigated, opens the possibility that this enigmatic ancient relic may be more widely active and physiologically relevant than previously appreciated.

Tissue microbiomes through the lenses of DNA sequencing, fluorescence, and electron microscopy

To further examine and validate our findings through other techniques beyond single-cell transcriptomics, we used DNA sequencing, fluorescence, and electron microscopy of virus-like particles (VLPs) and bacterial-like particles (BLPs) extracted from 4 tissue types (pancreas, blood, lymph node, and lung) and 8 donors (**Figure 5.a**). We developed an assay (see Methods) which was used to gently dissociate tissues, eliminate cell- or capsid-free nucleic acids, and subsequently size select for VLPs and BLPs. We used negative controls for all experimental steps to eliminate contaminants for bulk DNA sequencing of these extracted particles (see Methods). Using nucleic acid staining followed by fluorescence microscopy of these particles from the lung and lymph nodes

(**SI Figure 18**), we observed high concentrations of VLPs and BLPs (~10^7 BLPs and ~10^8 VLPs per cm^3 of tissue). Electron microscopy of VLPs further showed familiar virus morphologies such as a hexagonal capsid (similar to many bacteriophages) and rod-shaped capsids (reminiscent of the tobacco mosaic virus or the rudivirus) in the pancreas (**Figure 5.b**). It also revealed several novel morphologies, for example one with interior nodules and fibrous protrusions residing in a lymph node. Moreover, DNA sequencing of these extracted particles further captured numerous sequences mapping to viruses, bacteria, and even some fungi as some have similar sizes to bacteria (**Figure 5.c**).

Despite significant methodological differences, 50+ genera and ~70% of orders **(SI Table 6, Figure 5.d)** found through DNA sequencing of extracted VLPs and BLPs were also detected in the single-cell transcriptomic dataset. Interestingly, EBV and HERV K were both identified in this validation cohort. EBV was found in all 8 donors, which is a higher prevalence than obtained via RNA sequencing (3/15 donors). Considering that single-cell RNA sequencing preferentially captures transcriptionally active organisms, one would expect to see higher EBV donor positivity rate using DNA sequencing of VLPs. The combination of these two methodologies could perhaps be used to decipher active versus inactive organisms to further make a distinction between residents and passengers of the human tissue microbiome.

Moreover, we identified the species that overlap between all four datasets: TS, HMP, Nejman *et al.* and the validation dataset (**SI Figure 20**). For example, we found *Rothia mucilaginosa* in the pancreas and several other organs using single-cell transcriptomics (TS dataset). This species, a member of the airway microbiome (HMP dataset), appears in pancreatic, lung and breast tumors (Nejman *et al.*). Consistent with its traces in these three datasets, we identify this species through DNA sequencing of the extracted BLPs in the pancreas, lung, blood and lymph nodes. Given these validation experiments, it is perhaps fair to assume that sequences found in single-cell transcriptomic data are not all cell-or capsid-free. Some perhaps originate from intact organisms that escape cellular barriers to reach healthy tissues and tumors.

Discussion

The transcriptional footprints of microbial life in human organs and cells raise several questions. For example, how could microbes escape cellular barriers established by epithelial and endothelial cells? Are there any long-term residents of human tissues in a commensal relationship with our own cells? And, can these microbial signatures be used as diagnostic markers for organ health? One likely route of microbial migration is from the

gut to other organs, as leaky gut epithelia together with the gut-liver axis have been proposed as a route of dissemination for various microbial metabolites from the gut into the systemic circulation[59,60]. The same route is likely to harbor microbial genomes, transcripts, and perhaps whole cells and capsids. In fact, viral migration across epithelial layers has been demonstrated *in vitro*[61]. However, as we have shown there are many other possible flow paths from other external-facing microbiomes such as the oral cavity and the urogenital tract to various internal organs.

Overall, there is growing evidence for the existence of microbial interactions with human cells at sites in the body that are typically inaccessible to sampling. While microbial sequences have already been shown as diagnostic markers for cancer detection[14], our understanding of their associations with human cells is largely lacking. It is possible that causal associations exist, and that the list of oncogenic microbes is much larger than previously appreciated. Our finding of significant overlaps between healthy tissue microbiomes and tumor microbiomes suggests that tumor microbiomes are at least partially sourced from the healthy tissue microbiomes. As such, mechanism-level understanding of microbe-host interactions in human tissues could further shed light on the highly variable nature of individual responses to cancer immunotherapies[62,63] and provide novel and targeted avenues for microbiome-based immune modulation. Moreover, given that infectious agents account for ~15% of all cancers[64], and that many pathogens can manipulate the host cell cycle[65,66] and predispose host cells to cancer[67], it is highly plausible that we have only scratched the surface of causal links between microbes and various pathologies.

With the availability of high-throughput single cell sequencing technologies today, we stand at a unique time in the history of microbiology, where the study of microbe-host interactions offers an exciting new frontier. These interactions can now be studied at the level of individual cells[69,70,71,72] not just *in vitro,* but as we have shown, in the very real context of organs - in the presence of the immune system, neighboring cells, nutrient fluxes and many other important factors that shape microbial life cycles.

Conclusion

In this study we provide the first atlas of the human tissue microbiome with cell type resolution, for which we have developed new computational and experimental pipelines. We discovered numerous species, some of which are uncharacterized and potentially novel, from three domains of life. We map out the possible flow routes of microbes from external-facing microbiomes to internal tissues and tumors and show that tumor

microbiomes share many common species with healthy tissue microbiomes. Moreover, we demonstrate that at the species level, tissue microbiomes are not only donor-specific but also tissue-specific. We further constructed a network of associations between human cell types and microbes, shedding light on the types of organisms that cells from different tissues encounter either by way of immune clearance or physical proximity to microbial niches and sites of infections. This type of analysis can reveal the interactions between various cell types and microbial organisms. In cases of intra-cellular infectious events such as EBV, host and pathogen transcriptomic landscapes may reveal novel pathways that underlie cellular susceptibility and resistance, which would have great clinical significance.

Methods

Tabula Sapiens dataset

The TS dataset containing single-cell RNA sequencing data from Smart-seq2 and 10X droplet-based experimental pipelines were acquired from the Tabula Sapiens consortium[19] (https://github.com/czbiohub/tabula-sapiens). This included raw fastqs for donors 1-12, and 14-16 as well as TS objects which contained cell type annotations for 400,000+ cells (donor 16 cell type annotations were not included). All analysis was performed on the TS 10X data, which represents the vast majority of the dataset. Fastq files from empty wells acquired through Smart-seq2 experimental pipeline were used for one of our decontamination steps.

Bioinformatic pipeline: SIMBA

SIMBA is written in Snakemake[73] and consists of three connected workflows used for viral, bacterial and fungal sequence identification in single-cell human transcriptomic data. For preprocessing steps, we used UMI-tools[74] to filter and extract cell barcodes and UMIs from raw TS fastqs using the *extract* and *whitelist* commands. We used STAR[75] (v2.7) to align resulting sequences against a reference library containing the human genome (GRCh38.p13) and ERCCs. We then took unmapped sequences for further processing. We wrote *poly.py* to eliminate low-complexity sequences with high fractions of consecutive bases, often representing primer dimers and other sequences with long stretches of polyA.

To search for viral sequences, BLASTn[23] (v2.10) was used against the viral RefSeq database[76] (release 98) which included 9,000+ viral genomes. Any BLASTn alignment with E-values $<10^{-5}$ was considered a "significant hit". For each sequence the top significant hit (lowest E-value) was selected for downstream analysis. We wrote *dedup.py* to create input fasta files with sequences that produced a top hit in this step and use the resulting file to BLASTn against the nt database. The nt database at the time of download (11/02/2020) consisted of ~70 million sequences pooled from GenBank, EMBL, DDBJ, PDB, and RefSeq databases.

To search for bacterial and fungal sequences, BLASTn is used against microbeDB, a lean human microbiome-centric database we constructed for bacterial and fungal sequences (see Creating custom databases). Sequences that aligned to this database were selected and were used as input to BLASTn against slimBacteria_RefSeq, a reduced version of the bacterial RefSeq database[76] which we created to minimize computational time. Finally, taking the top hits from this step, we selected sequences that were input to BLASTn against the nt dataset. For sequences that did not produce significant hits, we used BLASTn against slimFungal_RefSeq, a reduced version of the fungal RefSeq database[76]. Finally, we selected top hits and used BLASTn against the nt database. The resulting top hits that emerged from the viral, bacterial, and fungal workflows against nt were combined into one dataset consisting of ~19 million sequences for further processing and quality control. All dependencies, commands, and parameters as well as computational resource requirements (CPU, memory and time) for scheduling Slurm job submissions are available for download (see Data and Materials availability).

Bioinformatic pipeline: taxonomic, alignment length, and percent identity filters

Taxonomic lineages for each top hit sequence were added using TaxonKit[77]. Top hits were filtered based on their taxonomic assignment to include only viral, bacterial, and fungal species. Only ~12% of the ~19 million sequences survived this filter. This percentage highlights the critical importance of our approach in having multiple checkpoints (first using STAR, then BLASTn against nt) to remove human and other non-microbial sequences. This further highlights the need for the computationally costly validation step against the nt database, and the insufficiency of relying on significant hits against custom microbial databases. Moreover, we selected only the highest-quality alignments against the nt database by applying fairly conservative alignment length (90 bp) and percent identity (90%) thresholds, and the resulting E-values were $< 10^{-21}$.

Bioinformatic pipeline: contamination filter

For bulk DNA sequencing of tissues, we used multiple negative controls for each step of the experimental workflow, including tissue digestion, DNA extraction, and library

preparation. We then created a list of contaminants and excluded those species if they were detected in the true samples. We further excluded species from more than 80 genera previously identified in commonly used reagents[78]. We also eliminated species that were found in empty wells (~100 wells) from the Smart-seq2 portion of the TS dataset. This was done because common to Smart-seq2 and 10X experimental pipelines are the tissue sample provider, lab space, some of the reagents, and the sequencing center.

In addition to these filters, to identify putative contaminants from TS dataset (10X portion), we employed a "prevalence" decontamination model similar to what has been proposed previously in the context of bulk metagenomic sequencing experiments[79]. The prevalence model is based on the premise that non-contaminants are enriched in true samples compared to negative control samples, whereas contaminants appear with similar frequency in negative control samples compared to true samples. This type of a model has the added benefit of being applicable to low microbial-biomass settings[79]. In our 10X-based search for microbial sequences in human tissues, empty droplets (those without a host cell) can be viewed as negative control samples, while cell-containing droplets can be viewed as true samples. Hence, we considered any microbial species with relative occurrence $\frac{c_{dts} / C_{dt}}{e_{dts} / E_{dt}} \geq R_{dts}$as a putative contaminant, where c_{dts} and e_{dts} represent the number of cell-containing droplets (*c*) and empty droplets (*e*) per donor (*d*) and tissue (*t*) sample that contain a given species (*s*). Moreover, C_{dt} and E_{dt} correspond to the total number of cell-containing droplets and empty droplets per donor (*d*) and tissue (*t*) sample. We conservatively chose $R_{dts} = 100$, to select only species that appear with more than 100 fold enrichment in cell-containing droplets compared to empty droplets within each sample. Counts of empty droplets were performed using Cell Ranger (v6.0.1) in addition to a custom script for downstream analysis (see Data and Materials availability). Together with the implementation of all filters and quality control measures, only ~0.5% of the original 19 million significant hits were retained for further analysis.

Bioinformatic pipeline: Creating custom databases

RefSeq databases[80] for bacterial and fungal genomes are available through NCBI, however they are typically large databases that would create significant computational bottlenecks. We created several custom databases to serve as lean human-microbiome centric databases for the initial iteration of BLASTn search for bacterial and fungal sequences. This was not done for viruses as they have much smaller genomes relative to bacteria and fungi and thus the viral RefSeq could be used directly.

MicrobeDB was constructed by combining two other custom databases, bacteriaDB and mycoDB. BacteriaDB was designed by taking the intersection of the HumGut database[81],

a comprehensive database of ~30,000 prokaryotic genomes mined from ~5,700 healthy human metagenomes, and the bacterial RefSeq database (downloaded on 01/2021). We therefore identified the unique set of species in HumGut and eliminated entries with unidentified species-level taxonomy. Similarly, we identified the unique set of bacterial RefSeq entries, and took the intersection of these two sets to arrive at several hundred bacterial species. We used a similar strategy to bias the initial search for fungi towards species previously identified in human mycobiome studies. We took five representative species from each genera found in both the fungal RefSeq (downloaded on 01/2021) and a recent survey of the healthy human mycobiome[82] to create mycoDB. While microbeDB was used for the first search iteration, larger databases were used for the second iteration. These larger databases, slimBacterial_RefSeq and slimFungal_RefSeq are the reduced versions of the bacterial and fungal RefSeq databases which we subsampled to include representative species and strains.

Bioinformatic pipeline: external databases

Tumor microbiome dataset was obtained from Nejman *et al.* (Table S4 and Figure 4D for tumor-enriched species) and only species and orders appearing in tumors that passed all of this study's contamination filters were selected for analysis. HMP dataset acquired from https://www.hmpdacc.org consisted of ~3000 entries primarily from bacterial strains from various isolation sites. We chose species found in external-facing microbiomes which form the majority of isolation sites (oral cavity, gastrointestinal tract, urogenital tract, airways, and the skin). When comparing the TS dataset to these two datasets, entries with unknown taxonomic assignments at the level of species or orders were excluded. Data on pathogenicity of different species was acquired from the HMP consortium[83] which was originally downloaded from the PATRIC database[40]. Additionally, we acquired antibacterial resistance profiles from the PATRIC database. We eliminated entries mapping to "intermediate resistance" or "not defined".

Tissue digestion and DNA extraction for the validation cohort

Tissues were acquired from Donor Network West (delivered on ice) and cut into $1cm^3$ blocks using UV and ethanol sterilized scalpels and petri dishes. Tissues were incubated for 90 minutes with collagenase IV (1 mg/mL, Worthington) at 37℃ and further disrupted by pipetting. Disrupted tissue samples were then incubated for 30 minutes at 37℃ with 19U/mL DNase I and 1mg/mL RNase A (ThermoFisher) - added to eliminate as much of the cell- or capsid-free DNA and RNA that is released during tissue digestion. The RNAase and DNAase enzymes were deactivated using RNAsecure at 1X and EDTA at

5mM for 10 minutes at 65℃. Each sample was then filtered through a sterile 40um cell strainer, to eliminate the cellular debris. The filtrate was passed through a 5um pore size syringe filter to enrich for particles in the size range of most microbes and to eliminate host cells. To separate out as many virus-like particles from this mixture of filtrate, we imposed another filter, this time using a 0.8um pore size syringe filter. The filter was inverted and objects larger than 0.8 um were collected. Although this was a protocol developed in-house, we were able to get key insight from several sources: Qiagen TissueLyser Handbook, Worthington Tissue Dissociation Guide and viral metagenomic studies[84]. DNA extraction was then performed using QIAmp DNA mini kit (Qiagen) according to kit instructions.

TEM.

We plasma treated EM grids (300 mesh Carbon/Formvar coated Copper grids) using Leica EM ACE600 to allow each sample to bind to the grid. We applied 5 uL of 1:1000 diluted sample onto each EM grid and allowed the sample to bind for 3 minutes. The grid was then exposed to 1% uranyl acetate for one minute, and subsequently dried using a filter paper. Samples were imaged using JEOL JEM-1400.

Fluorescence imaging.

Following tissue digestion and microbial extraction, filterates were stained with SYBR™ Gold Nucleic Acid Gel Stain (ThermoFisher) at 10X for 10 minutes at room temperature and imaged using Leica CTR6000. Bright dots were counted manually across different sections across a hemocytometer grid and average counts were projected to estimate VLP and BLP concentrations per cubic centimeter of tissue.

Citations

1. Gilbert, J. *et al.* Current understanding of the human microbiome. *Nat. Med.* **24**, 392–400 (2018).
2. Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, 61–66 (2017).
3. Kowarsky, M. *et al.* Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA. *Proc. Natl. Acad. Sci.* **114**,

9623–9628 (2017).

4. Mahmoudabadi, G. & Phillips, R. A comprehensive and quantitative exploration of thousands of viral genomes. *eLife* **7**, e31955 (2018).
5. Dickson, R. P., Erb-Downward, J. R., Martinez, F. J. & Huffnagle, G. B. The Microbiome and the Respiratory Tract. *Annu. Rev. Physiol.* **78**, 481–504 (2016).
6. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457**, 480–484 (2009).
7. Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60 (2012).
8. Cryan, J. F., O'Riordan, K. J., Sandhu, K., Peterson, V. & Dinan, T. G. The gut microbiome in neurological disorders. *Lancet Neurol.* **19**, 179–194 (2020).
9. Sharon, G. *et al.* Human Gut Microbiota from Autism Spectrum Disorder Promote Behavioral Symptoms in Mice. *Cell* **177**, 1600-1618.e17 (2019).
10. Asnicar, F. *et al.* Microbiome connections with host metabolism and habitual diet from 1,098 deeply phenotyped individuals. *Nat. Med.* **27**, 321–332 (2021).
11. Jie, Z. *et al.* The gut microbiome in atherosclerotic cardiovascular disease. *Nat. Commun.* **8**, 845 (2017).
12. Fujimura, K. E. *et al.* Neonatal gut microbiota associates with childhood multisensitized atopy and T cell differentiation. *Nat. Med.* **22**, 1187–1191 (2016).
13. Wolter, M. *et al.* Leveraging diet to engineer the gut microbiome. *Nat. Rev. Gastroenterol. Hepatol.* 1–18 (2021) doi:10.1038/s41575-021-00512-7.
14. Poore, G. D. *et al.* Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* **579**, 567–574 (2020).

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

16. Pushalkar, S. *et al.* The Pancreatic Cancer Microbiome Promotes Oncogenesis by Induction of Innate and Adaptive Immune Suppression. *Cancer Discov.* **8**, 403–416 (2018).

17. Riquelme, E. *et al.* Tumor Microbiome Diversity and Composition Influence Pancreatic Cancer Outcomes. *Cell* **178**, 795-806.e12 (2019).

18. Eslami-S, Z., Majidzadeh-A, K., Halvaei, S., Babapirali, F. & Esmaeili, R. Microbiome and Breast Cancer: New Role for an Ancient Population. *Front. Oncol.* **10**, 120 (2020).

19. Consortium, T. T. S. & Quake, S. R. *The Tabula Sapiens: a single cell transcriptomic atlas of multiple organs from individual human donors*. 2021.07.19.452956 https://www.biorxiv.org/content/10.1101/2021.07.19.452956v1 (2021) doi:10.1101/2021.07.19.452956.

20. van Tilburg Bernardes, E. *et al.* Intestinal fungi are causally implicated in microbiome assembly and immune development in mice. *Nat. Commun.* **11**, 2577 (2020).

21. Forbes, J. D., Bernstein, C. N., Tremlett, H., Van Domselaar, G. & Knox, N. C. A Fungal World: Could the Gut Mycobiome Be Involved in Neurological Disease? *Front. Microbiol.* **9**, 3249 (2019).

22. Shkoporov, A. N. *et al.* The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. *Cell Host Microbe* **26**, 527-541.e5 (2019).

23. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local

alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

24. Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46 (2014).

25. Woellmer, A. & Hammerschmidt, W. Epstein-Barr virus and host cell methylation: Regulation of latency, replication and virus reactivation. *Curr. Opin. Virol.* **3**, 260–265 (2013).

26. Tierney, R. J., Shannon-Lowe, C. D., Fitzsimmons, L., Bell, A. I. & Rowe, M. Unexpected patterns of Epstein–Barr virus transcription revealed by a High throughput PCR array for absolute quantification of viral mRNA. *Virology* **474**, 117–130 (2015).

27. Qin, J. *et al.* A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).

28. Al Bataineh, M. T. *et al.* Revealing links between gut microbiome and its fungal community in Type 2 Diabetes Mellitus among Emirati subjects: A pilot study. *Sci. Rep.* **10**, 9624 (2020).

29. Milo, R. & Phillips, R. *Cell Biology by the numbers*. (Garland Science, 2015). doi:10.1201/9780429258770.

30. Schaechter, M., MaalØe, O. & Kjeldgaard, N. O. Y. 1958. Dependency on Medium and Temperature of Cell Size and Chemical Composition during Balanced Growth of Salmonella typhimurium. *Microbiology* **19**, 592–606.

31. Scott, M., Gunderson, C. W., Mateescu, E. M., Zhang, Z. & Hwa, T. Interdependence of cell growth and gene expression: origins and consequences. *Science* **330**, 1099–1102 (2010).

32. Milo, R., Jorgensen, P., Moran, U., Weber, G. & Springer, M. BioNumbers--the

database of key numbers in molecular and cell biology. *Nucleic Acids Res.* **38**, D750-753 (2010).

33. Human Microbiome Project Consortium. A framework for human microbiome research. *Nature* **486**, 215–221 (2012).

34. Chukkapalli, S. S. *et al.* Invasion of Oral and Aortic Tissues by Oral Spirochete Treponema denticola in ApoE−/− Mice Causally Links Periodontal Disease and Atherosclerosis. *Infect. Immun.* **82**, 1959–1967 (2014).

35. Santos, P. B. do R. E. dos *et al.* Review- The periodontal pathogen Treponema denticola: an atherosclerosis risk factor. *Res. Soc. Dev.* **10**, e25810111637–e25810111637 (2021).

36. Okuda, K. *et al.* Detection of Treponema denticola in Atherosclerotic Lesions. *J. Clin. Microbiol.* (2001) doi:10.1128/JCM.39.3.1114-1117.2001.

37. *Streptococcus pyogenes : Basic Biology to Clinical Manifestations*. (University of Oklahoma Health Sciences Center, 2016).

38. Sampaio-Maia, B., Caldas, I. M., Pereira, M. L., Pérez-Mongiovi, D. & Araujo, R. Chapter Four - The Oral Microbiome in Health and Its Implication in Oral and Systemic Diseases. in *Advances in Applied Microbiology* (eds. Sariaslani, S. & Michael Gadd, G.) vol. 97 171–210 (Academic Press, 2016).

39. Schenkein, H. A., Papapanou, P. N., Genco, R. & Sanz, M. Mechanisms underlying the association between periodontitis and atherosclerotic disease. *Periodontol. 2000* **83**, 90–106 (2020).

40. Davis, J. J. *et al.* The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. *Nucleic Acids Res.* **48**, D606–D612 (2020).

41. Franzosa, E. A. *et al.* Identifying personal microbiomes using metagenomic codes. *Proc. Natl. Acad. Sci.* (2015) doi:10.1073/pnas.1423854112.

42. Integrative HMP (iHMP) Research Network Consortium. The Integrative Human Microbiome Project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell Host Microbe* **16**, 276–289 (2014).

43. Kapono, C. A. *et al.* Creating a 3D microbial and chemical snapshot of a human habitat. *Sci. Rep.* **8**, 3669 (2018).

44. Aykut, B. *et al.* The fungal mycobiome promotes pancreatic oncogenesis via activation of MBL. *Nature* **574**, 264–267 (2019).

45. Young, L. S., Yap, L. F. & Murray, P. G. Epstein–Barr virus: more than 50 years old and still providing surprises. *Nat. Rev. Cancer* **16**, 789–802 (2016).

46. Dunmire, S. K., Verghese, P. S. & Balfour, H. H. Primary Epstein-Barr virus infection. *J. Clin. Virol.* **102**, 84–92 (2018).

47. Thorley-Lawson, D. A. EBV Persistence—Introducing the Virus. in *Epstein Barr Virus Volume 1: One Herpes Virus: Many Diseases* (ed. Münz, C.) 151–209 (Springer International Publishing, 2015). doi:10.1007/978-3-319-22822-8_8.

48. Smatti, M. K. *et al.* Epstein–Barr Virus Epidemiology, Serology, and Genetic Variability of LMP-1 Oncogene Among Healthy Population: An Update. *Front. Oncol.* **8**, 211 (2018).

49. Bjornevik, K. *et al.* Longitudinal analysis reveals high prevalence of Epstein-Barr virus associated with multiple sclerosis. *Science* (2022) doi:10.1126/science.abj8222.

50. Narita, Y. *et al.* Pin1 Interacts with the Epstein-Barr Virus DNA Polymerase Catalytic Subunit and Regulates Viral DNA Replication. *J. Virol.* **87**, 2120–2127 (2013).

51. van Zyl, D. G., Mautner, J. & Delecluse, H.-J. Progress in EBV Vaccines. *Front. Oncol.* **9**, 104 (2019).

52. Chiu, S.-H. *et al.* Epstein-Barr Virus BALF3 Has Nuclease Activity and Mediates Mature Virion Production during the Lytic Cycle. *J. Virol.* **88**, 4962–4975 (2014).

53. Wen, W. *et al.* Epstein-Barr Virus BZLF1 Gene, a Switch from Latency to Lytic Infection, Is Expressed as an Immediate-Early Gene after Primary Infection of B Lymphocytes. *J. Virol.* **81**, 1037–1042 (2007).

54. Dervan, E., Bhattacharyya, D. D., McAuliffe, J. D., Khan, F. H. & Glynn, S. A. Ancient Adversary – HERV-K (HML-2) in Cancer. *Front. Oncol.* **11**, 1707 (2021).

55. Agoni, L., Guha, C. & Lenz, J. Detection of Human Endogenous Retrovirus K (HERV-K) Transcripts in Human Prostate Cancer Cell Lines. *Front. Oncol.* **3**, 180 (2013).

56. Chan, S. M. *et al.* The HERV-K accessory protein Np9 controls viability and migration of teratocarcinoma cells. *PLOS ONE* **14**, e0212970 (2019).

57. Grow, E. J. *et al.* Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **522**, 221–225 (2015).

58. Wang, T. *et al.* Regulation of stem cell function and neuronal differentiation by HERV-K via mTOR pathway. *Proc. Natl. Acad. Sci.* **117**, 17842–17853 (2020).

59. Adawi, D., Molin, G. & Jeppsson, B. Gut–liver axis. *HPB* **1**, 173–186 (1999).

60. Tripathi, A. *et al.* The gut–liver axis and the intersection with the microbiome. *Nat. Rev. Gastroenterol. Hepatol.* **15**, 397–411 (2018).

61. Nguyen, S. *et al.* Bacteriophage Transcytosis Provides a Mechanism To Cross Epithelial Cell Layers. *mBio* **8**, e01874-17.

62. Zhou, C.-B., Zhou, Y.-L. & Fang, J.-Y. Gut Microbiota in Cancer Immune Response and Immunotherapy. *Trends Cancer* **7**, 647–660 (2021).

63. York, A. Gut microbiota sways response to cancer immunotherapy. *Nat. Rev. Microbiol.* **16**, 121–121 (2018).

64. Martel, C. de, Georges, D., Bray, F., Ferlay, J. & Clifford, G. M. Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis. *Lancet Glob. Health* **8**, e180–e190 (2020).

65. Evans, E. L., Becker, J. T., Fricke, S. L., Patel, K. & Sherer, N. M. HIV-1 Vif's Capacity To Manipulate the Cell Cycle Is Species Specific. *J. Virol.* **92**, e02102-17.

66. Fan, Y., Sanyal, S. & Bruzzone, R. Breaking Bad: How Viruses Subvert the Cell Cycle. *Front. Cell. Infect. Microbiol.* **8**, 396 (2018).

67. Gagnaire, A., Nadel, B., Raoult, D., Neefjes, J. & Gorvel, J.-P. Collateral damage: insights into bacterial mechanisms that predispose host cells to cancer. *Nat. Rev. Microbiol.* **15**, 109–128 (2017).

68. Duan, Y. *et al.* Bacteriophage targeting of gut bacterium attenuates alcoholic liver disease. *Nature* **575**, 505–511 (2019).

69. Russell, A. B., Trapnell, C. & Bloom, J. D. Extreme heterogeneity of influenza virus infection in single cells. *eLife* **7**, e32303 (2018).

70. Zanini, F., Pu, S.-Y., Bekerman, E., Einav, S. & Quake, S. R. Single-cell transcriptional dynamics of flavivirus infection. *eLife* **7**, e32942 (2018).

71. Sun, J. *et al.* Single cell heterogeneity in influenza A virus gene expression shapes the innate antiviral response to infection. *PLOS Pathog.* **16**, e1008671 (2020).

72. Xue, Y. *et al.* A single-parasite transcriptional atlas of Toxoplasma Gondii reveals

novel control of antigen expression. *eLife* **9**, e54129 (2020).

73. Mölder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Research* **10**, 33 (2021).

74. Smith, T. S., Heger, A. & Sudbery, I. UMI-tools: Modelling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.* gr.209601.116 (2017) doi:10.1101/gr.209601.116.

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

76. Brister, J. R., Ako-Adjei, D., Bao, Y. & Blinkova, O. NCBI viral genomes resource. *Nucleic Acids Res.* **43**, D571-577 (2015).

77. Shen, W. & Ren, H. TaxonKit: A practical and efficient NCBI taxonomy toolkit. *J. Genet. Genomics* **48**, 844–850 (2021).

78. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12**, 87 (2014).

79. 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).

80. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733-745 (2016).

81. Hiseni, P., Rudi, K., Wilson, R. C., Hegge, F. T. & Snipen, L. HumGut: a comprehensive human gut prokaryotic genomes collection filtered by metagenome data. *Microbiome* **9**, 165 (2021).

82. Tiew, P. Y. *et al.* The Mycobiome in Health and Disease: Emerging Concepts, Methodologies and Challenges. *Mycopathologia* **185**, 207–231 (2020).

83. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).

84. Weynberg, K. D., Wood-Charlson, E. M., Suttle, C. A. & van Oppen, M. J. H. Generating viral metagenomes from the coral holobiont. *Front. Microbiol.* **5**, 206 (2014).

Acknowledgements

We thank Rob Phillips for insightful comments on the manuscript. We also thank Leila Sidow, Robert Jones, Roja Kolluru, Ed Marti, Elizabeth Jerrison, Mark Kowarsky, Alina Isakova, Sheela Crasta, Yuan Xue, Mira Moufarrej, Michael Swift, and all the other members of the Quake lab, along with the Tabula Sapiens and the Boundaries of Life consortia for their help and suggestions towards experimental and computational pipelines. We are grateful to Donor Network West personnel including Ahmad Salehi and Ravi Ponnusamy, as well as the Cell Sciences Imaging Facility at Stanford University. We are especially indebted to donors and their families for their incredible contribution to science. This work was supported by the Chan Zuckerberg Biohub, the Gordon and Betty Moore Foundation and the John Templeton Foundation.

Funding

This work was supported by the Chan Zuckerberg Biohub, the Gordon and Betty Moore Foundation and the John Templeton Foundation.

Author contributions

Conceptualization: GM, SRQ

Methodology: GM, SRQ
Investigation: GM, SRQ
Data Curation: GM
Formal Analysis: GM
Visualization: GM
Software: GM
Funding acquisition: GM, Tabula Sapiens Consortium, SRQ
Resources: GM, Tabula Sapiens Consortium, SRQ
Project administration: GM, Tabula Sapiens Consortium, SRQ
Writing – original draft: GM
Writing – review & editing: GM, SRQ

Competing interests

Authors declare that they have no competing interests.

The Tabula Sapiens Consortium Author List

Overall Project Direction and Coordination

Robert C. Jones[1] , Jim Karkanias[2] , Mark Krasnow[3,4], Angela Oliveira Pisco[2] , Stephen R. Quake[1,2,5], Julia Salzman[3,6], Nir Yosef[2,7,8,9]

Donor Recruitment

Bryan Bulthaup[10], Phillip Brown[10], William Harper[10], Marisa Hemenez[10], Ravikumar Ponnusamy[10], Ahmad Salehi[10], Bhavani A. Sanagavarapu[10], Eileen Spallino[10]

Surgeons

Ksenia A. Aaron[11], Waldo Concepcion[10], James M. Gardner[12,13], Burnett Kelly[10,14], Nikole Neidlinger[10], Zifa Wang[10]

Logistical coordination

Sheela Crasta[1,2], Saroja Kolluru[1,2], Maurizio Morri2 , Angela Oliveira Pisco[2] , Serena Y. Tan[15], Kyle J. Travaglini[3] , Chenling Xu[7]

Organ Processing

Marcela Alcántara-Hernández[16], Nicole Almanzar[17], Jane Antony[18], Benjamin Beyersdorf[19], Deviana Burhan[20], Kruti Calcuttawala[21], Matthew M. Carter[16], Charles K. F. Chan[18,22], Charles A. Chang[23], Stephen Chang[3,19], Alex Colville[21,24], Sheela Crasta[1,2], Rebecca N. Culver[25], Ivana Cvijović[1,5], Gaetano D'Amato[26], Camille Ezran[3] , Francisco X. Galdos[18], Astrid Gillich[3] , William R. Goodyer[27], Yan Hang[23,28], Alyssa Hayashi[1] , Sahar Houshdaran[29], Xianxi Huang[19,30], Juan C. Irwin[29], SoRi Jang[3] , Julia Vallve Juanico[29], Aaron M. Kershner[18], Soochi Kim[21,24], Bernhard Kiss[18], Saroja Kolluru[1,2], William Kong[18], Maya E. Kumar[17], Angera H. Kuo[18], Rebecca Leylek[16], Baoxiang Li[31], Gabriel B. Loeb[32], Wan-Jin Lu[18], Sruthi Mantri[33], Maxim Markovic[1] , Patrick L. McAlpine[11,34], Antoine de Morree[21,24], Maurizio Morri[2] , Karim Mrouj[18], Shravani Mukherjee[31], Tyler Muser[17], Patrick Neuhöfer[3,35,36], Thi D. Nguyen[37], Kimberly Perez[16], Ragini Phansalkar[26], Angela Oliveira Pisco[2] , Nazan Puluca[18] , Zhen Qi18, Poorvi Rao[20], Hayley Raquer-McKay[16], Nicholas Schaum[18,21], Bronwyn Scott[31], Bobak Seddighzadeh[38], Joe Segal[20], Sushmita Sen[29], Shaheen Sikandar[18], Sean P. Spencer[16], Lea Steffes[17], Varun R. Subramaniam[31], Aditi Swarup[31], Michael Swift[1], Kyle J. Travaglini[3] , Will Van Treuren[16], Emily Trimm[26], Stefan Veizades[19,39], Sivakamasundari Vijayakumar[18], Kim Chi Vo[29], Sevahn K. Vorperian[1,40], Wanxin Wang[29], Hannah N.W. Weinstein[38], Juliane Winkler[41], Timothy T.H. Wu[3] , Jamie Xie[38], Andrea R.Yung[3] , Yue Zhang[3]

Sequencing

Angela M. Detweiler[2] , Honey Mekonen[2] , Norma F. Neff[2] , Rene V. Sit[2] , Michelle Tan[2] , Jia Yan[2]

Histology

Gregory R. Bean[15], Vivek Charu[15], Erna Forgó[15], Brock A. Martin[15], Michael G. Ozawa[15], Oscar Silva[15], Serena Y. Tan[15], Angus Toland[15], Venkata N.P. Vemuri[2]

Data Analysis

Shaked Afik[7] , Kyle Awayan[2] , Rob Bierman[3,6], Olga Borisovna Botvinnik[2] , Ashley Byrne[2] , Michelle Chen[1] , Roozbeh Dehghannasiri[3,6], Angela M. Detweiler[2] , Adam Gayoso[7] , Alejandro A Granados[2] , Qiqing Li[2] , Gita Mahmoudabadi[1] , Aaron McGeever[2] , Antoine de Morree[21,24], Julia Eve Olivieri[3,6,42], Madeline Park[2] , Angela Oliveira Pisco[2] , Neha Ravikumar[1] , Julia Salzman[3,6], Geoff Stanley[1] , Michael Swift[1] , Michelle Tan[2] , Weilun Tan[2] , Alexander J Tarashansky[2] , Rohan Vanheusden[2] , Sevahn K. Vorperian[1,40], Peter Wang[3,6], Sheng Wang[2] , Galen Xing[2] , Chenling Xu[6] , Nir Yosef[2,6,7,8]

Expert Cell Type Annotation

Marcela Alcántara-Hernández[16], Jane Antony[18], Charles K. F. Chan[18,22], Charles A. Chang[23], Alex Colville[21,24], Sheela Crasta[1,2], Rebecca Culver[25], Les Dethlefsen[43], Camille Ezran[3] , Astrid Gillich[3] , Yan Hang[23,28], Po-Yi Ho[16], Juan C. Irwin[29], SoRi Jang[3] , Aaron M. Kershner[18], William Kong[18], Maya E Kumar[17], Angera H. Kuo[18], Rebecca Leylek[16], Shixuan Liu[3,44], Gabriel B. Loeb[32], Wan-Jin Lu[18], Jonathan S Maltzman[45,46], Ross J. Metzger[27,47], Antoine de Morree[21,24], Patrick Neuhöfer[3,35,36], Kimberly Perez[16], Ragini Phansalkar[26], Zhen Qi[18], Poorvi Rao[20], Hayley RaquerMcKay[16], Koki Sasagawa[19], Bronwyn Scott[31], Rahul Sinha[15,18,35], Hanbing Song[38], Sean P. Spencer[16], Aditi Swarup[31], Michael Swift[1] , Kyle J. Travaglini[3] , Emily Trimm[26], Stefan Veizades[19,39], Sivakamasundari Vijayakumar[18], Bruce Wang[20], Wanxin Wang[29], Juliane Winkler[41], Jamie Xie[38], Andrea R.Yung[3]

Tissue Expert Principal Investigators

Steven E. Artandi[3,35,36], Philip A. Beachy[18,23,48], Michael F. Clarke[18], Linda C. Giudice[29], Franklin W. Huang[38,49], Kerwyn Casey Huang[1,16], Juliana Idoyaga[16], Seung K Kim[23,28], Mark Krasnow[3,4], Christin S. Kuo[17], Patricia Nguyen[19,39,46], Stephen R. Quake[1,2,5], Thomas A. Rando[21,24], Kristy Red-Horse[26], Jeremy Reiter[50], David A. Relman[16,43,46], Justin L. Sonnenburg[16], Bruce Wang[20], Albert Wu[31], Sean M. Wu[19,39], Tony Wyss-Coray[21,24]

Affiliations

[1]Department of Bioengineering, Stanford University; Stanford, CA, USA.
[2]Chan Zuckerberg Biohub; San Francisco, CA, USA.
[3]Department of Biochemistry, Stanford University School of Medicine; Stanford, CA, USA.
[4]Howard Hughes Medical Institute; USA.
[5]Department of Applied Physics, Stanford University; Stanford, CA, USA.
[6]Department of Biomedical Data Science, Stanford University; Stanford, CA, USA.
[7]Center for Computational Biology, University of California Berkeley; Berkeley, CA, USA.
[8]Department of Electrical Engineering and Computer Sciences, University of California Berkeley; Berkeley, CA, USA.
[9]Ragon Institute of MGH, MIT and Harvard; Cambridge, MA, USA.
[10]Donor Network West; San Ramon, CA, USA.
[11]Department of Otolaryngology-Head and Neck Surgery, Stanford University School of Medicine; Stanford, California, USA.

[12]Department of Surgery, University of California San Francisco; San Francisco, CA, USA. [13]Diabetes Center, University of California San Francisco; San Francisco, CA, USA.
[14]DCI Donor Services; Sacramento, CA, USA.
[15]Department of Pathology, Stanford University School of Medicine; Stanford, CA, USA.
[16]Department of Microbiology and Immunology, Stanford University School of Medicine; Stanford, CA, USA.
[17]Department of Pediatrics, Division of Pulmonary Medicine, Stanford University; Stanford, CA, USA.
[18]Institute for Stem Cell Biology and Regenerative Medicine, Stanford University School of Medicine; Stanford, CA, USA.
[19]Department of Medicine, Division of Cardiovascular Medicine, Stanford University; Stanford, CA, USA.
[20]Department of Medicine and Liver Center, University of California San Francisco; San Francisco, CA, USA.
[21]Department of Neurology and Neurological Sciences, Stanford University School of Medicine; Stanford, CA, USA.
[22]Department of Surgery - Plastic and Reconstructive Surgery, Stanford University School of Medicine; Stanford, CA, USA.
[23]Department of Developmental Biology, Stanford University School of Medicine, Stanford, CA, USA
[24]Paul F. Glenn Center for the Biology of Aging, Stanford University School of Medicine; Stanford, CA, USA.
[25]Department of Genetics, Stanford University School of Medicine; Stanford, CA, USA.
[26]Department of Biology, Stanford University; Stanford, CA, USA.
[27]Department of Pediatrics, Division of Cardiology, Stanford University School of Medicine; Stanford, CA, USA.
[28]Stanford Diabetes Research Center, Stanford University School of Medicine, Stanford, California
[29]Center for Gynecology and Reproductive Sciences, Department of Obstetrics, Gynecology and Reproductive Sciences, University of California San Francisco; San Francisco, CA, USA. [30]Department of Critical Care Medicine, The First Affiliated Hospital of Shantou University Medical College; Shantou, China.
[31]Department of Ophthalmology, Stanford University School of Medicine; Stanford, CA, USA. [32]Division of Nephrology, Department of Medicine, University of California San Francisco; San Francisco, CA, USA.
[33]Stanford University School of Medicine; Stanford, CA, USA.
[34]Mass Spectrometry Platform, Chan Zuckerberg Biohub; Stanford, CA, USA.

[35]Stanford Cancer Institute, Stanford University School of Medicine; Stanford, CA, USA.
[36]Department of Medicine, Division of Hematology, Stanford University School of Medicine, Stanford, CA, USA
[37]Department of Biochemistry and Biophysics, Cardiovascular Research Institute, University of California San Francisco; San Francisco, CA, USA.
[38]Division of Hematology and Oncology, Department of Medicine, Bakar Computational Health Sciences Institute, Institute for Human Genetics, University of California San Francisco; San Francisco, CA, USA.
[39]Stanford Cardiovascular Institute; Stanford CA, USA.
[40]Department of Chemical Engineering, Stanford University; Stanford, CA, USA.
[41]Department of Cell & Tissue Biology, University of California San Francisco; San Francisco, CA, USA.
[42]Institute for Computational and Mathematical Engineering, Stanford University; Stanford, CA, USA.
[43]Division of Infectious Diseases & Geographic Medicine, Department of Medicine, Stanford University School of Medicine; Stanford, CA, USA.
[44]Department of Chemical and Systems Biology, Stanford University School of Medicine, Stanford, CA, USA
[45]Division of Nephrology, Stanford University School of Medicine; Stanford, CA, USA.
[46]Veterans Affairs Palo Alto Health Care System; Palo Alto, CA, USA.
[47]Vera Moulton Wall Center for Pulmonary and Vascular Disease, Stanford University School of Medicine; Stanford, CA, USA.
[48]Department of Urology, Stanford University School of Medicine, Stanford, CA, USA
[49]Division of Hematology/Oncology, Department of Medicine, San Francisco Veterans Affairs Health Care System, San Francisco, CA, USA.
[50]Department of Biochemistry, University of California San Francisco; San Francisco, CA, USA.

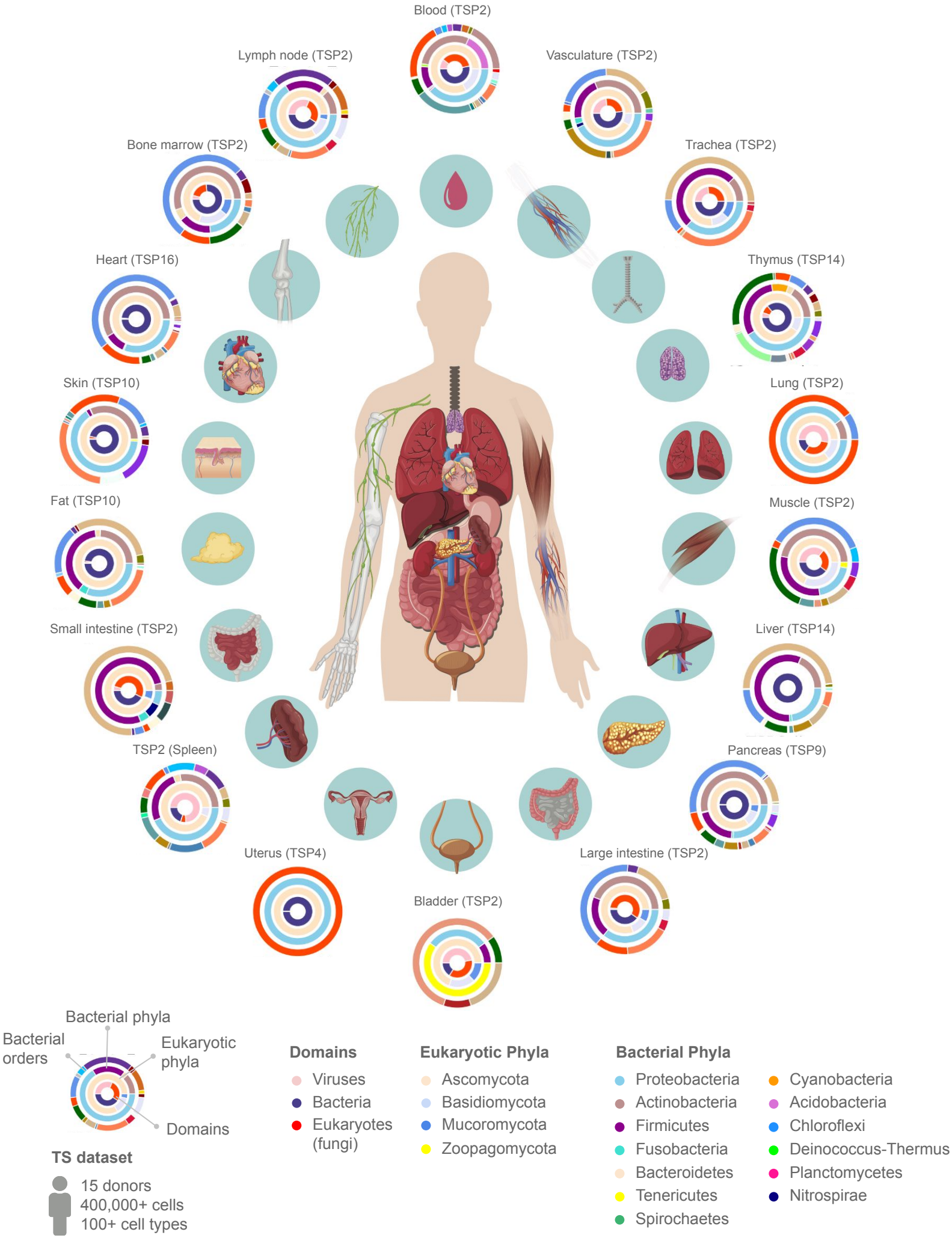
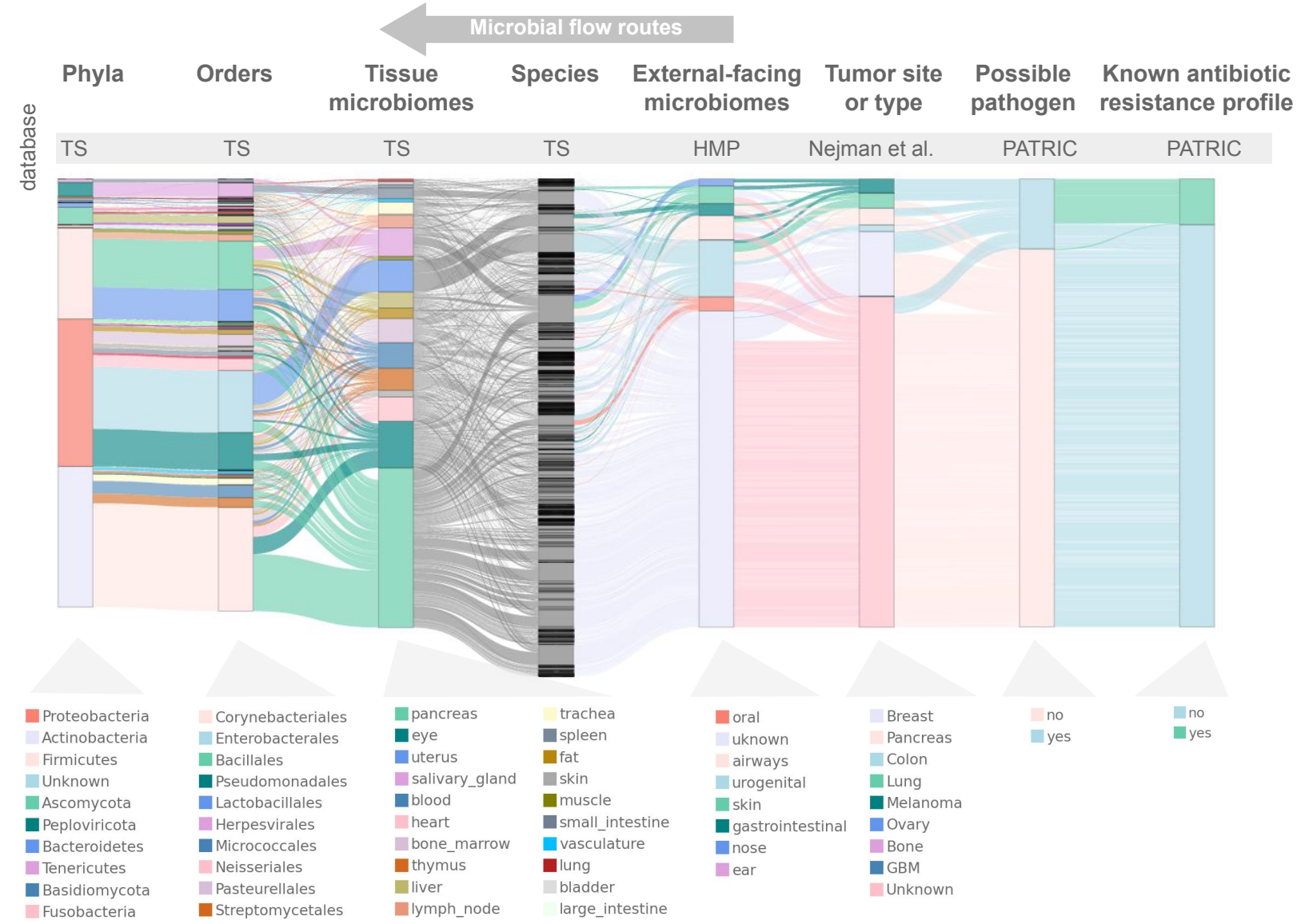
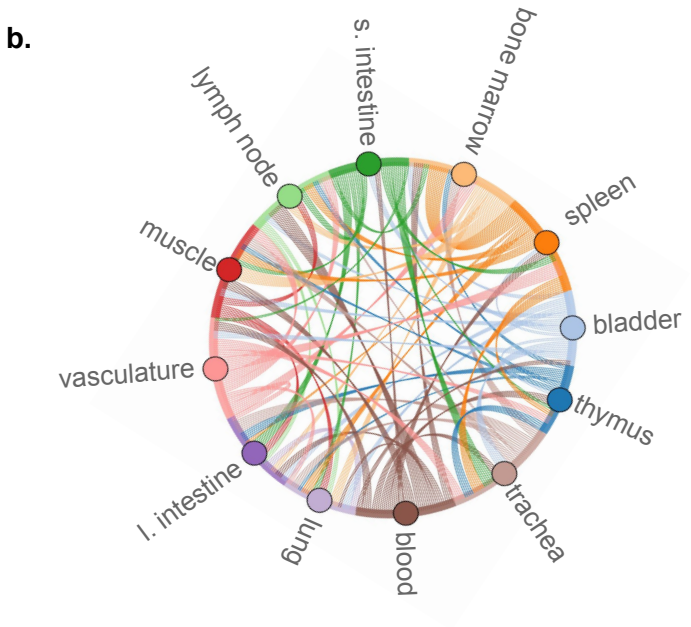

Blood (TSP2)
Lymph node (TSP2)
Vasculature (TSP2)
Bone marrow (TSP2)
Trachea (TSP2)
Heart (TSP16)
Thymus (TSP14)
Skin (TSP10)
Lung (TSP2)
Fat (TSP10)
Muscle (TSP2)
Small intestine (TSP2)
Liver (TSP14)
TSP2 (Spleen)
Pancreas (TSP9)
Uterus (TSP4)
Large intestine (TSP2)
Bladder (TSP2)
Bacterial phyla
Bacterial orders
Eukaryotic phyla
Domains
TS dataset
15 donors
400,000+ cells
100+ cell types
Domains
Viruses
Bacteria
Eukaryotes (fungi)
Eukaryotic Phyla
Ascomycota
Basidiomycota
Mucoromycota
Zoopagomycota
Bacterial Phyla
Proteobacteria
Actinobacteria
Firmicutes
Fusobacteria
Bacteroidetes
Tenericutes
Spirochaetes
Cyanobacteria
Acidobacteria
Chloroflexi
Deinococcus-Thermus
Planctomycetes
Nitrospirae

Figure 1. Tissue microbiome atlas built from the Tabula Sapiens (TS) single-cell transcriptomic dataset consisting of 20 tissue types from 15 donors (identified by TSP#), 100+ cell types and 400,000+ annotated cells. Tissue microbiomes are stratified based on taxonomy: domains (inner ring) followed by eukaryotic phyla, bacterial phyla and bacterial orders (legend shown in SI Figure 2), respectively. Sequences are grouped based on donor (one donor is shown per tissue, see SI Figure 2 for the complete set of donors for each tissue type).


Microbial flow routes
Phyla
Orders
Tissue microbiomes
Species
External-facing microbiomes
Tumor site or type
Possible pathogen
Known antibiotic resistance profile
database
TS
TS
TS
TS
HMP
Nejman et al.
PATRIC
PATRIC
Proteobacteria
Actinobacteria
Firmicutes
Unknown
Ascomycota
Peploviricota
Bacteroidetes
Tenericutes
Basidiomycota
Fusobacteria
Corynebacteriales
Enterobacterales
Bacillales
Pseudomonadales
Lactobacillales
Herpesvirales
Micrococcales
Neisseriales
Pasteurellales
Streptomycetales
pancreas
eye
uterus
salivary_gland
blood
heart
bone_marrow
thymus
liver
lymph_node
trachea
spleen
fat
skin
muscle
small_intestine
vasculature
lung
bladder
large_intestine
oral
uknown
airways
urogenital
skin
gastrointestinal
nose
ear
Breast
Pancreas
Colon
Lung
Melanoma
Ovary
Bone
GBM
Unknown
no
yes
no
yes


b.
s. intestine
bone marrow
lymph node
spleen
muscle
bladder
vasculature
thymus
l. intestine
trachea
lung
blood

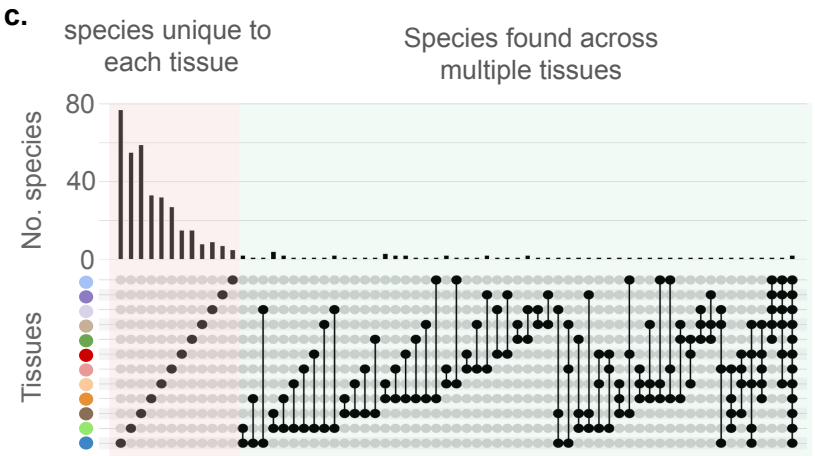

c.
species unique to each tissue
Species found across multiple tissues
No. species
80
40
0
Tissues

Figure 2. a) Sankey plot of ~30,000 surviving sequences, revealing the taxonomic classification and tissue of origin for each sequence (each line in the plot represents an individual sequence) along with additional data from the Human Microbiome Project (HMP) database, tumor microbiome dataset by Nejman *et al.*, and the PATRIC database. Sequences belonging to species that are not found in these additional databases, including all viral and fungal species, are labeled as “unknown” or “no” for the last four columns. Only the top ten most abundant phyla and orders are denoted in the legend. b) Circos plot of shared species within an individual donor 2 (TSP2). Each link represents a unique species. Species that are not shared between tissues are not shown. c) UpSet plot for donor 2 demonstrating the number of shared species between tissues, as well as the number of species unique to each tissue. Tissues are color-coded according to panel b.

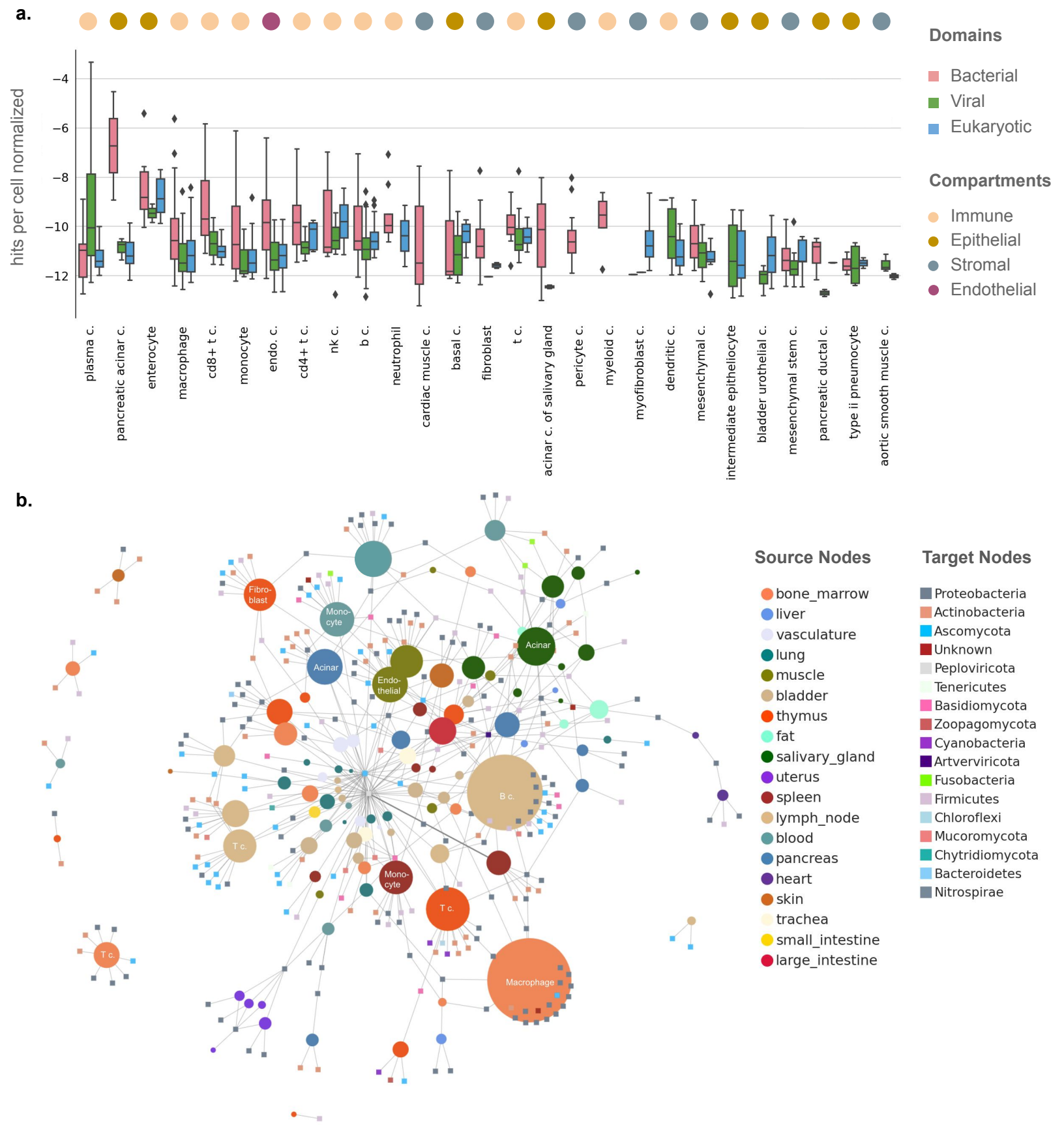
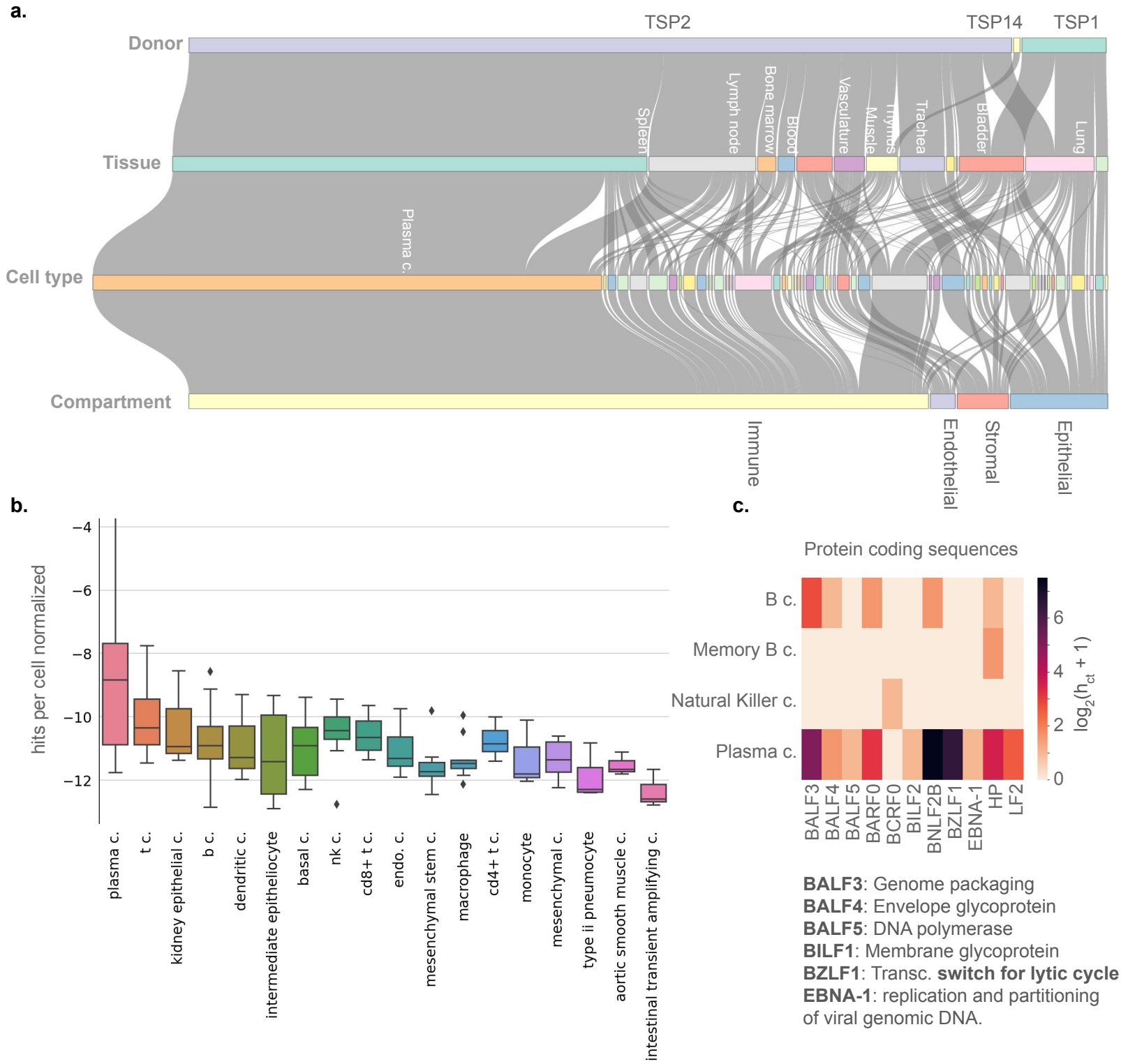

a.
hits per cell normalized
−4
−6
−8
−10
−12
plasma c.
pancreatic acinar c.
enterocyte
macrophage
cd8+ t c.
monocyte
endo. c.
cd4+ t c.
nk c.
b c.
neutrophil
cardiac muscle c.
basal c.
fibroblast
t c.
acinar c. of salivary gland
pericyte c.
myeloid c.
myofibroblast c.
dendritic c.
mesenchymal c.
intermediate epitheliocyte
bladder urothelial c.
mesenchymal stem c.
pancreatic ductal c.
type ii pneumocyte
aortic smooth muscle c.
Domains
Bacterial
Viral
Eukaryotic
Compartments
Immune
Epithelial
Stromal
Endothelial
b.
Fibro-
blast
Mono-
cyte
Acinar
Endo-
thelial
Acinar
B c.
T c.
Mono-
cyte
T c.
T c.
Macrophage
Source Nodes
bone_marrow
liver
vasculature
lung
muscle
bladder
thymus
fat
salivary_gland
uterus
spleen
lymph_node
blood
pancreas
heart
skin
trachea
small_intestine
large_intestine
Target Nodes
Proteobacteria
Actinobacteria
Ascomycota
Unknown
Peploviricota
Tenericutes
Basidiomycota
Zoopagomycota
Cyanobacteria
Artverviricota
Fusobacteria
Firmicutes
Chloroflexi
Mucoromycota
Chytridiomycota
Bacteroidetes
Nitrospirae

Figure 3. Cellular microbiomes across 15 donors. a) Distributions of hits per cell against three domains of life, grouped by cell type. Cell types are ordered based on the maximum value found in any of the the three distributions. Y-axis corresponds to $\log_2(h_c/g_c)$, where h_c is the number of hits per cell, normalized by the number of g_c, number of detected genes per cell. Only cells with detected hits are included in the analysis. b) Network of human cell types (grouped by tissue of origin) in association with microbes. Cell types are source nodes (circular), and microbial species are the target nodes (square). Source nodes are weighted by $S_i/\log_2 C_i$, where S_i represents the number of unique species associated with source node *i*, and is normalized by the total number of cells, C, sequenced for source node *i*. The normalization ensures that cell types with greater representation in the dataset do not artificially appear to have more diverse microbial associations. Moreover, edges in this network are weighted by $N_{j,i}$, the counts of sequences mapping to microbial species *j* associated with source node *i.* For visual clarity, only larger source nodes are labeled.



BALF3: Genome packaging
BALF4: Envelope glycoprotein
BALF5: DNA polymerase
BILF1: Membrane glycoprotein
BZLF1: Transc. **switch for lytic cycle**
EBNA-1: replication and partitioning of viral genomic DNA.

Figure 4. Detection of EBV in single-cell RNA seq. data from three donors. a) Sankey plot of EBV sequences, demonstrating their donor, tissue, cell type, and compartment of origin. For visual clarity, only large groups have been labeled for tissue and cell type rows. b) EBV hit per cell distributions grouped by cell type (TSP2 or donor 2). Y axis denotes $\log_2(h_c/g_c)$, where h_c is the number of hits per cell, normalized by the number of g_c, number of detected genes per cell. c) heatmap of protein-coding sequences identified in donor 2 mapping to EBV, grouped by cell type. The scale bar denotes $\log_2(h_{ct} + 1)$, where h_{ct} represents number of hits per cell type. The functional role of some of the EBV protein coding sequences are denoted. Note, "HP" corresponds to "hypothetical protein".

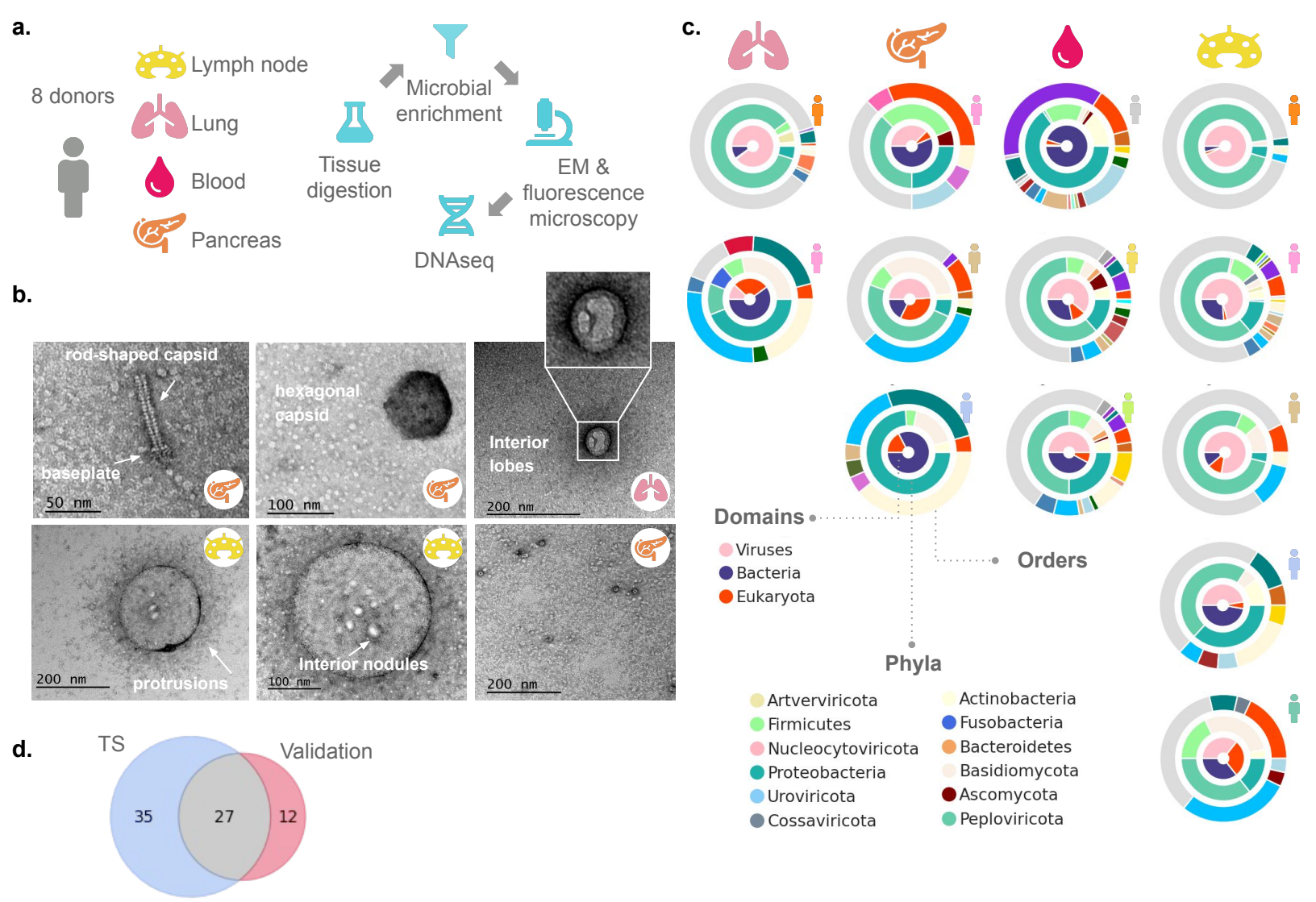
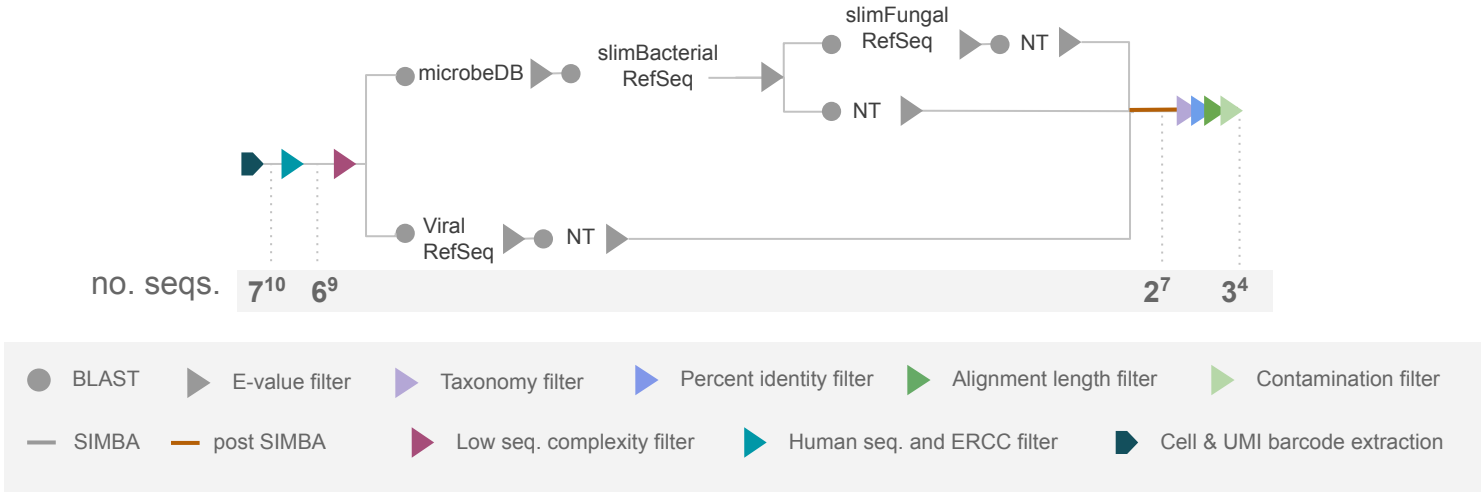

a.
8 donors
Lymph node
Lung
Blood
Pancreas
Microbial enrichment
Tissue digestion
EM & fluorescence microscopy
DNAseq
b.
rod-shaped capsid
baseplate
50 nm
hexagonal capsid
100 nm
Interior lobes
200 nm
200 nm
protrusions
Interior nodules
100 nm
200 nm
d.
TS
Validation
35
27
12
c.
Domains
Viruses
Bacteria
Eukaryota
Orders
Phyla
Artverviricota
Firmicutes
Nucleocytoviricota
Proteobacteria
Uroviricota
Cossaviricota
Actinobacteria
Fusobacteria
Bacteroidetes
Basidiomycota
Ascomycota
Peploviricota

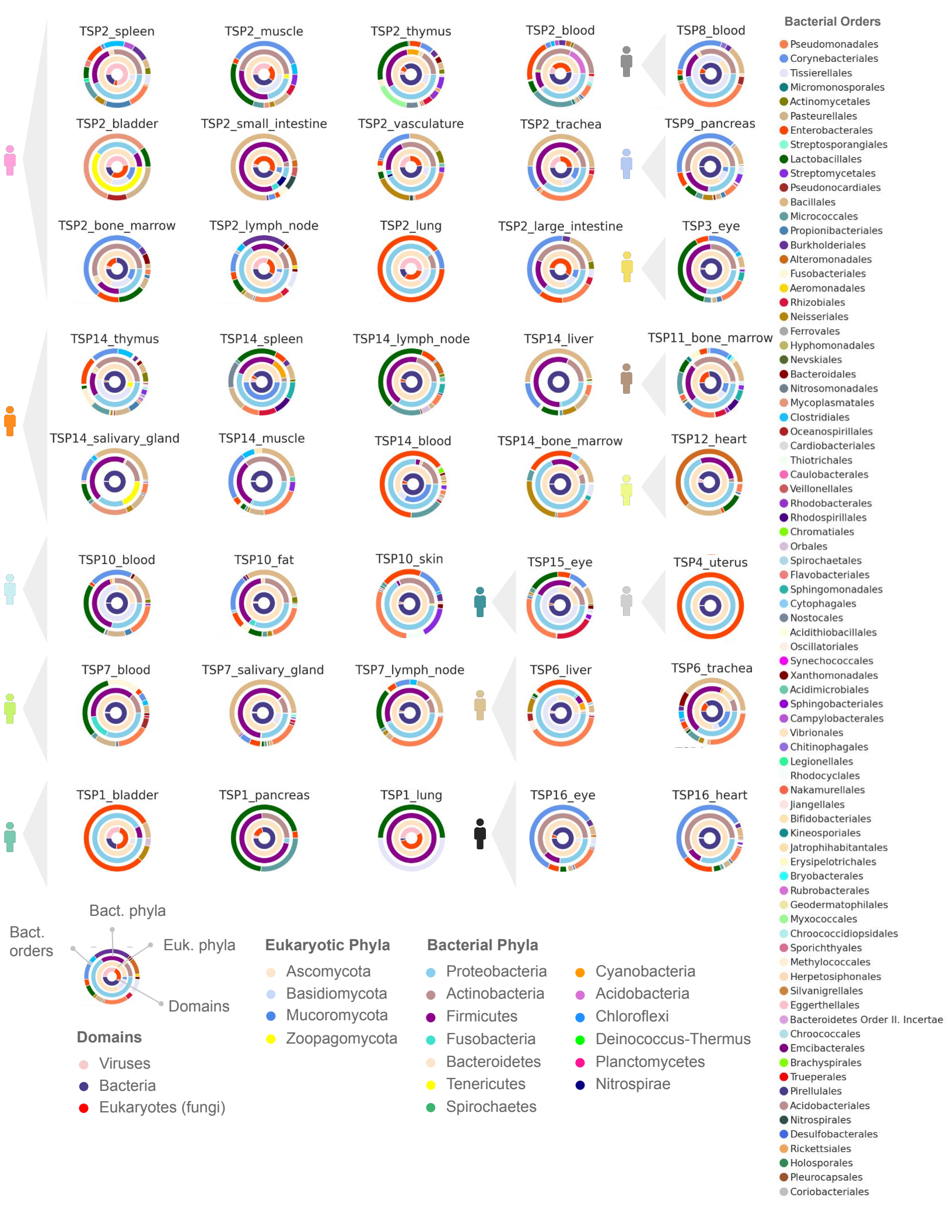
Figure 5. Tissue microbiomes from 8 donors using DNA sequencing, fluorescence and EM imaging. **a)** study design and experimental workflow. **b)** EM images of viral fractions retrieved from the pancreas, lung and lymph nodes showing both known and novel virus-like morphologies. **c)** BLAST hits per tissue and donor stratified based on taxonomy: domains (inner ring), phyla and orders (outer ring, legends are plotted in **SI Figure 19**). **d)** venn diagram of orders found in the TS and the validation datasets derived from the subset of tissue types shared between the two datasets.

SI

SI Figures + captions

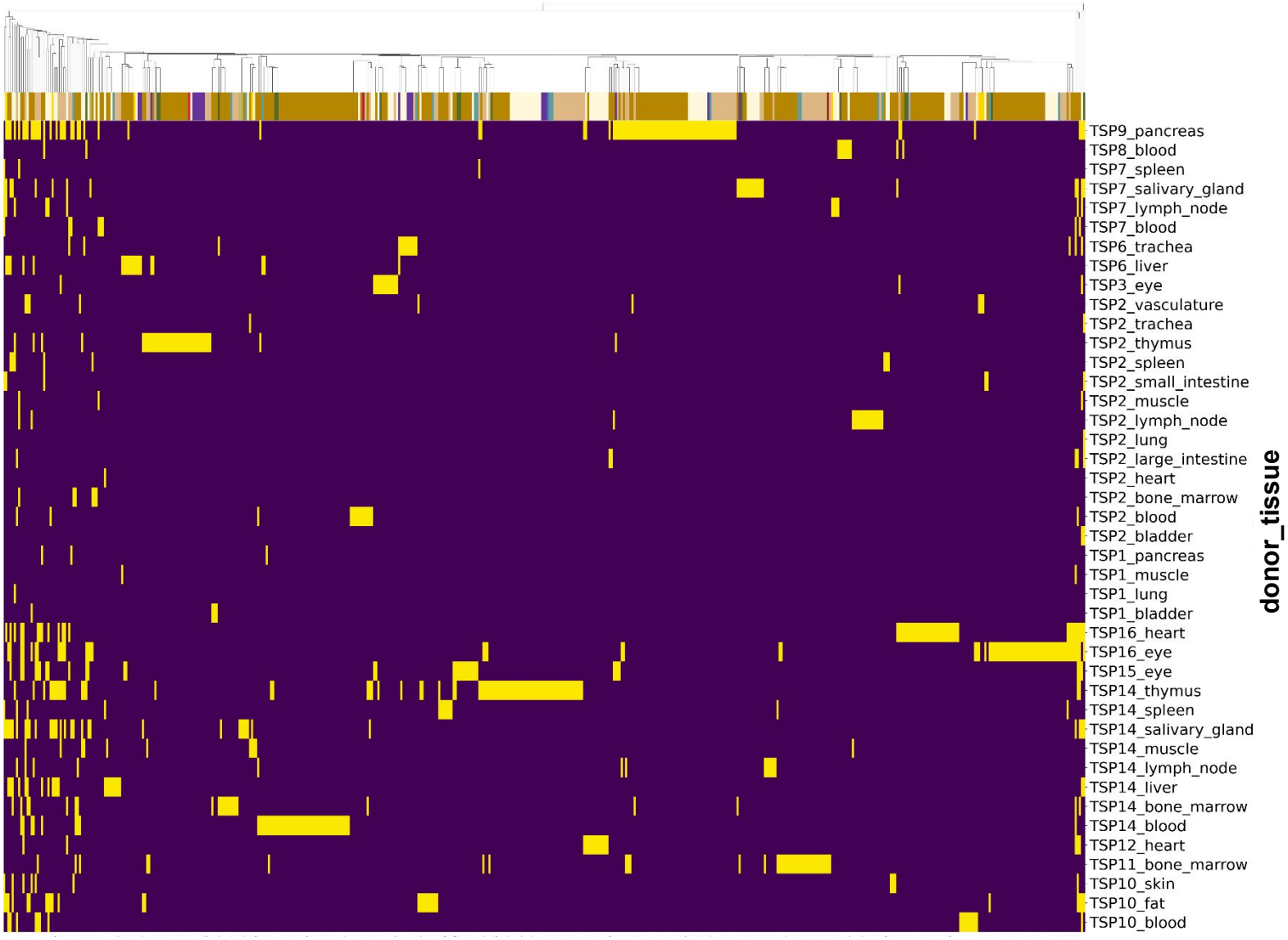

slimFungal
RefSeq
NT
microbeDB
slimBacterial
RefSeq
NT
Viral
RefSeq
NT
no. seqs.
7^{10}
6^{9}
2^{7}
3^{4}
BLAST
E-value filter
Taxonomy filter
Percent identity filter
Alignment length filter
Contamination filter
SIMBA
post SIMBA
Low seq. complexity filter
Human seq. and ERCC filter
Cell & UMI barcode extraction

SI Figure 1. Computational pipeline developed for the identification of high-confidence bacterial, viral and fungal sequences in single-cell data. Number of surviving sequences at different stages of the pipeline are shown (starting with ~70 billion sequences and arriving at ~30,000).


TSP2_spleen
TSP2_muscle
TSP2_thymus
TSP2_blood
TSP8_blood
TSP2_bladder
TSP2_small_intestine
TSP2_vasculature
TSP2_trachea
TSP9_pancreas
TSP2_bone_marrow
TSP2_lymph_node
TSP2_lung
TSP2_large_intestine
TSP3_eye
TSP14_thymus
TSP14_spleen
TSP14_lymph_node
TSP14_liver
TSP11_bone_marrow
TSP14_salivary_gland
TSP14_muscle
TSP14_blood
TSP14_bone_marrow
TSP12_heart
TSP10_blood
TSP10_fat
TSP10_skin
TSP15_eye
TSP4_uterus
TSP7_blood
TSP7_salivary_gland
TSP7_lymph_node
TSP6_liver
TSP6_trachea
TSP1_bladder
TSP1_pancreas
TSP1_lung
TSP16_eye
TSP16_heart
Bacterial Orders
Pseudomonadales
Corynebacteriales
Tissierellales
Micromonosporales
Actinomycetales
Pasteurellales
Enterobacterales
Streptosporangiales
Lactobacillales
Streptomycetales
Pseudonocardiales
Bacillales
Micrococcales
Propionibacteriales
Burkholderiales
Alteromonadales
Fusobacteriales
Aeromonadales
Rhizobiales
Neisseriales
Ferrovales
Hyphomonadales
Nevskiales
Bacteroidales
Nitrosomonadales
Mycoplasmatales
Clostridiales
Oceanospirillales
Cardiobacteriales
Thiotrichales
Caulobacterales
Veillonellales
Rhodobacterales
Rhodospirillales
Chromatiales
Orbales
Spirochaetales
Flavobacteriales
Sphingomonadales
Cytophagales
Nostocales
Acidithiobacillales
Oscillatoriales
Synechococcales
Xanthomonadales
Acidimicrobiales
Sphingobacteriales
Campylobacterales
Vibrionales
Chitinophagales
Legionellales
Rhodocyclales
Nakamurellales
Jiangellales
Bifidobacteriales
Kineosporiales
Jatrophihabitantales
Erysipelotrichales
Bryobacterales
Rubrobacterales
Geodermatophilales
Myxococcales
Chroococcidiopsidales
Sporichthyales
Methylococcales
Herpetosiphonales
Silvanigrellales
Eggerthellales
Bacteroidetes Order II. Incertae
Chroococcales
Emcibacterales
Brachyspirales
Trueperales
Pirellulales
Acidobacteriales
Nitrospirales
Desulfobacterales
Rickettsiales
Holosporales
Pleurocapsales
Coriobacteriales
Bact. phyla
Bact. orders
Euk. phyla
Domains
Domains
Viruses
Bacteria
Eukaryotes (fungi)
Eukaryotic Phyla
Ascomycota
Basidiomycota
Mucoromycota
Zoopagomycota
Bacterial Phyla
Proteobacteria
Actinobacteria
Firmicutes
Fusobacteria
Bacteroidetes
Tenericutes
Spirochaetes
Cyanobacteria
Acidobacteria
Chloroflexi
Deinococcus-Thermus
Planctomycetes
Nitrospirae

SI Figure 2. Tissue microbiome atlas. Tissue microbiomes are stratified based on taxonomy: domains (inner ring) followed by eukaryotic phyla, bacterial phyla and bacterial orders, respectively. Sequences are grouped based on donor and tissue.

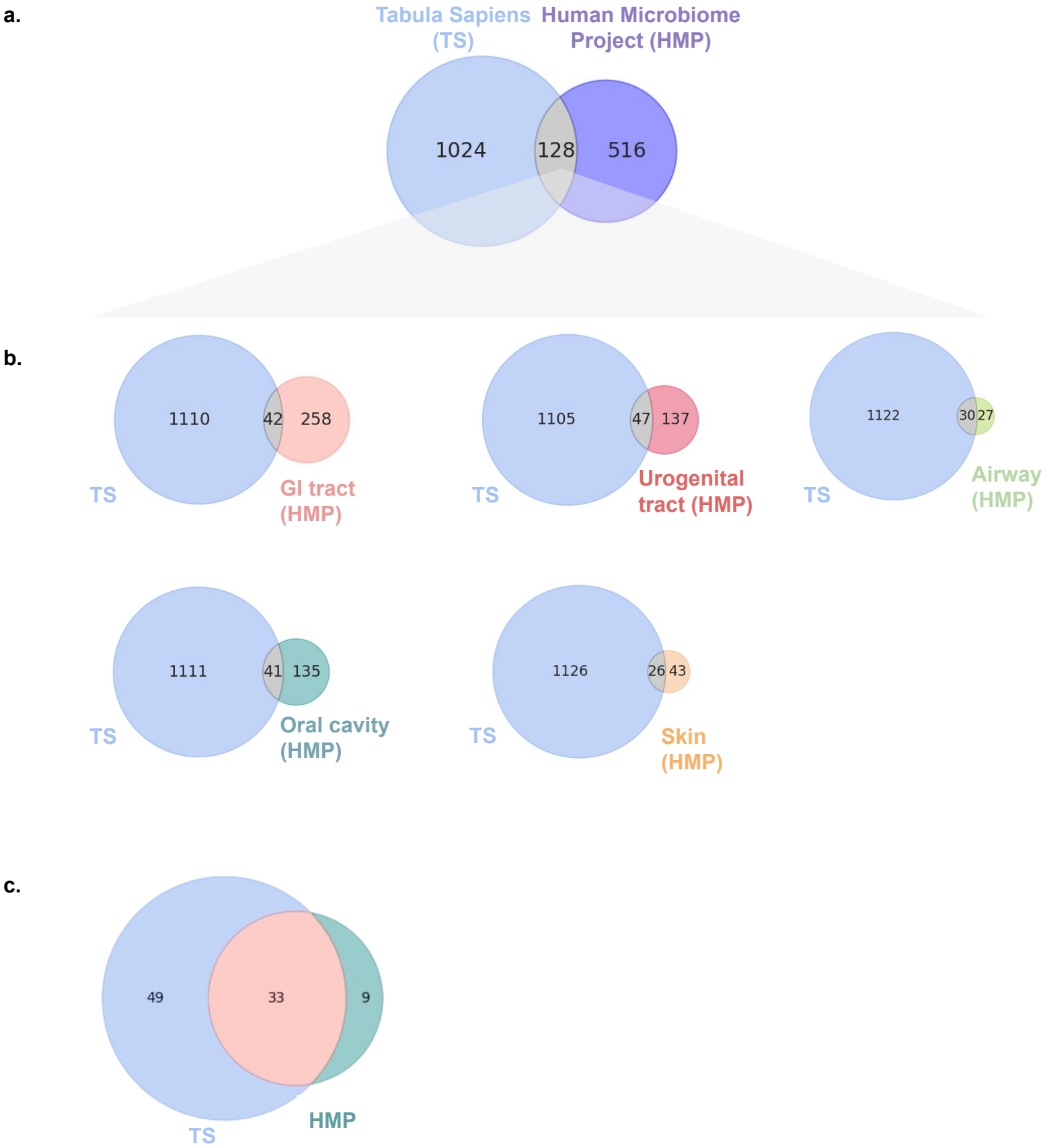
Bacterial species (based on ribosomal RNA) colored by phyla



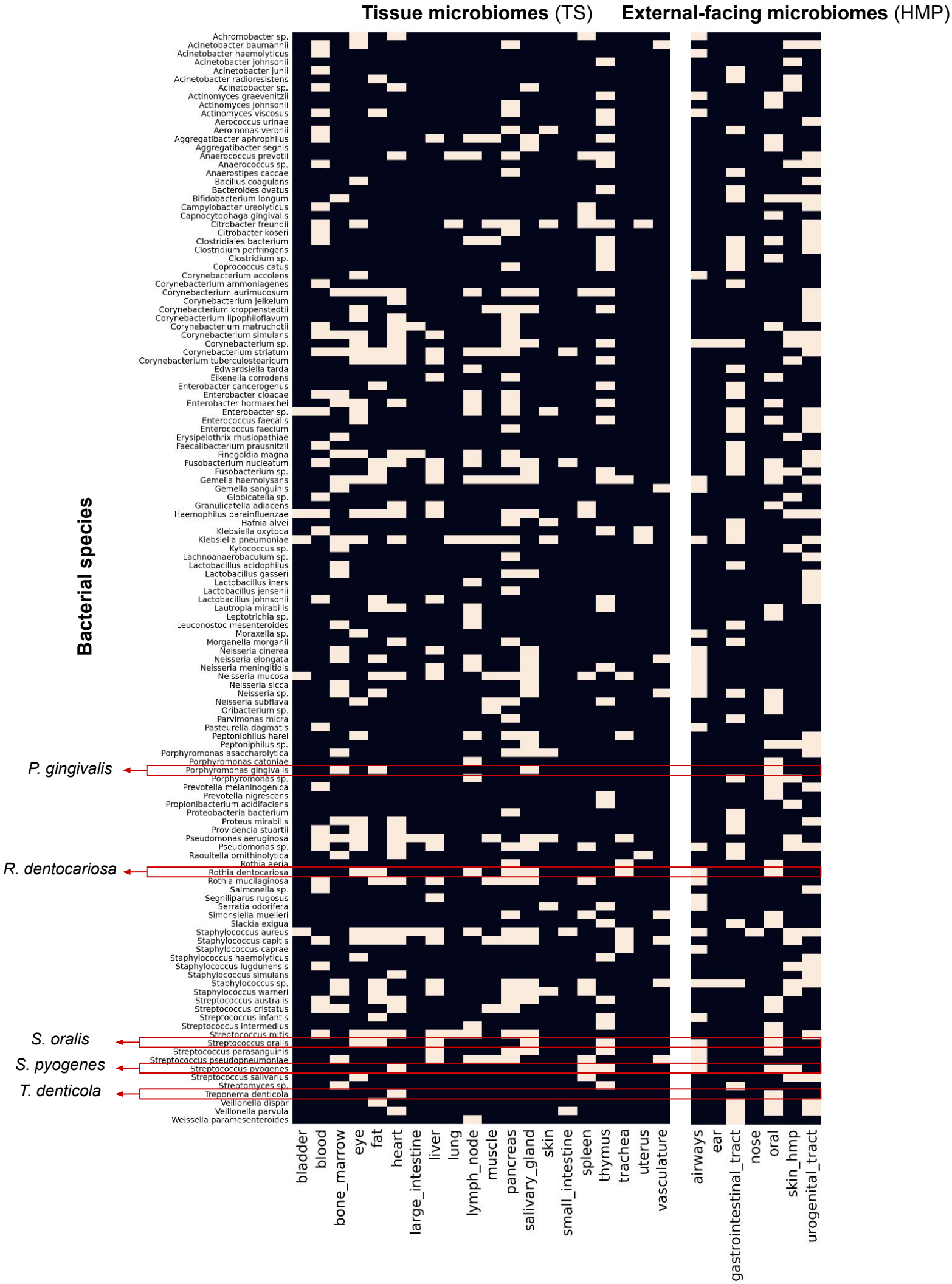
Bacterial phyla

- Acidobacteria
- Actinobacteria
- Bacteroidetes
- Chloroflexi
- Cyanobacteria
- Deinococcus-Thermus
- Firmicutes
- Fusobacteria
- Nitrospirae
- Proteobacteria
- Spirochaetes
- Tenericutes
- Unknown

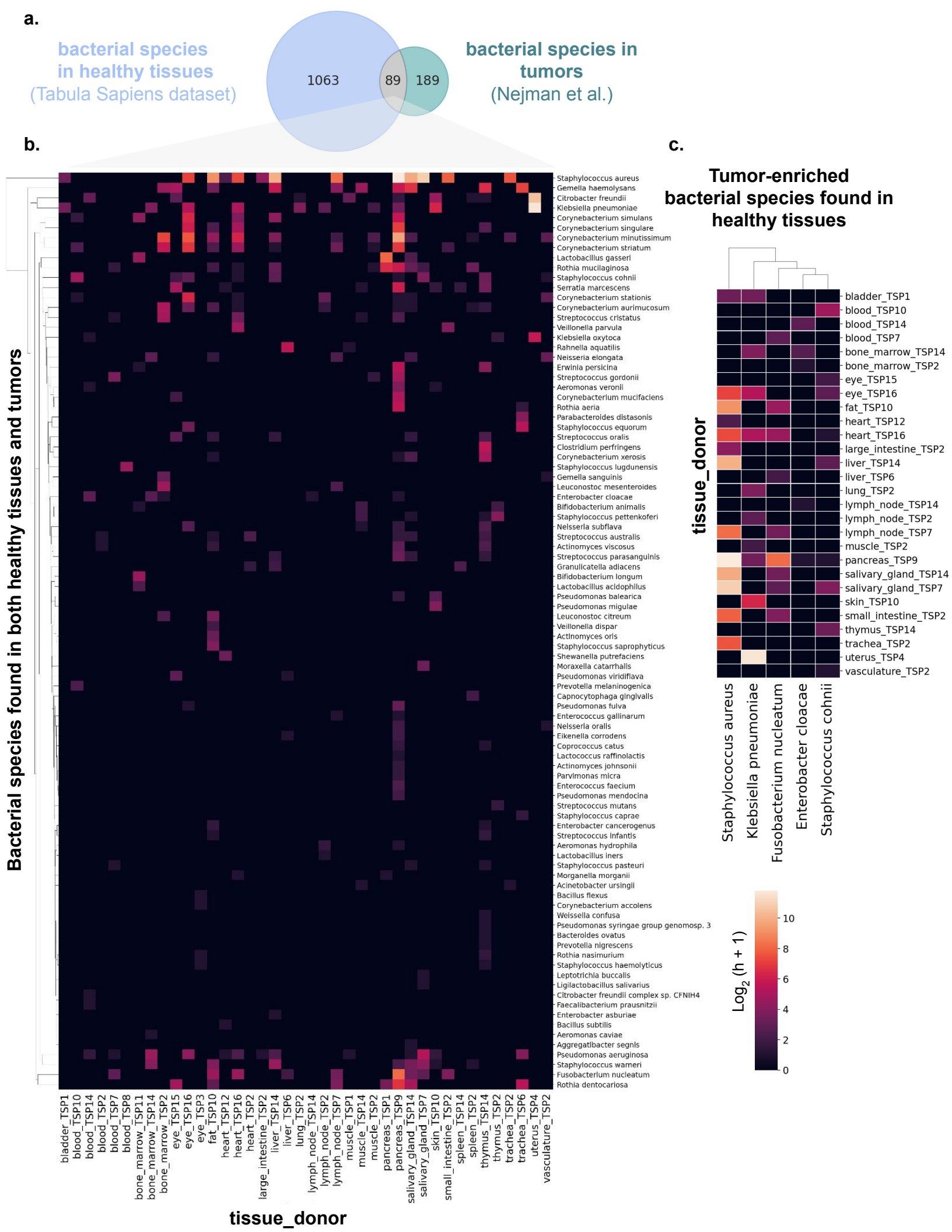
SI Figure 3. Ribosomal RNA from different bacterial phyla across various tissue types. Clustered heatmap of ribosomal hits (presence/yellow and absence/blue is shown) found in each donor's tissue sample. Bacterial species are color-coded based on their phylum-level classification.


a.
Tabula Sapiens (TS)
Human Microbiome Project (HMP)
1024
128
516
b.
1110
42
258
TS
GI tract (HMP)
1105
47
137
TS
Urogenital tract (HMP)
1122
30
27
TS
Airway (HMP)
1111
41
135
TS
Oral cavity (HMP)
1126
26
43
TS
Skin (HMP)
c.
49
33
9
TS
HMP

SI Figure 4. Comparisons between the Tabula Sapiens (TS) dataset and the Human Microbiome Project (HMP). a) Number of bacterial species shared between TS and HMP, b) broken down by HMP isolation sites. c) order-level comparison between TS and HMP datasets.

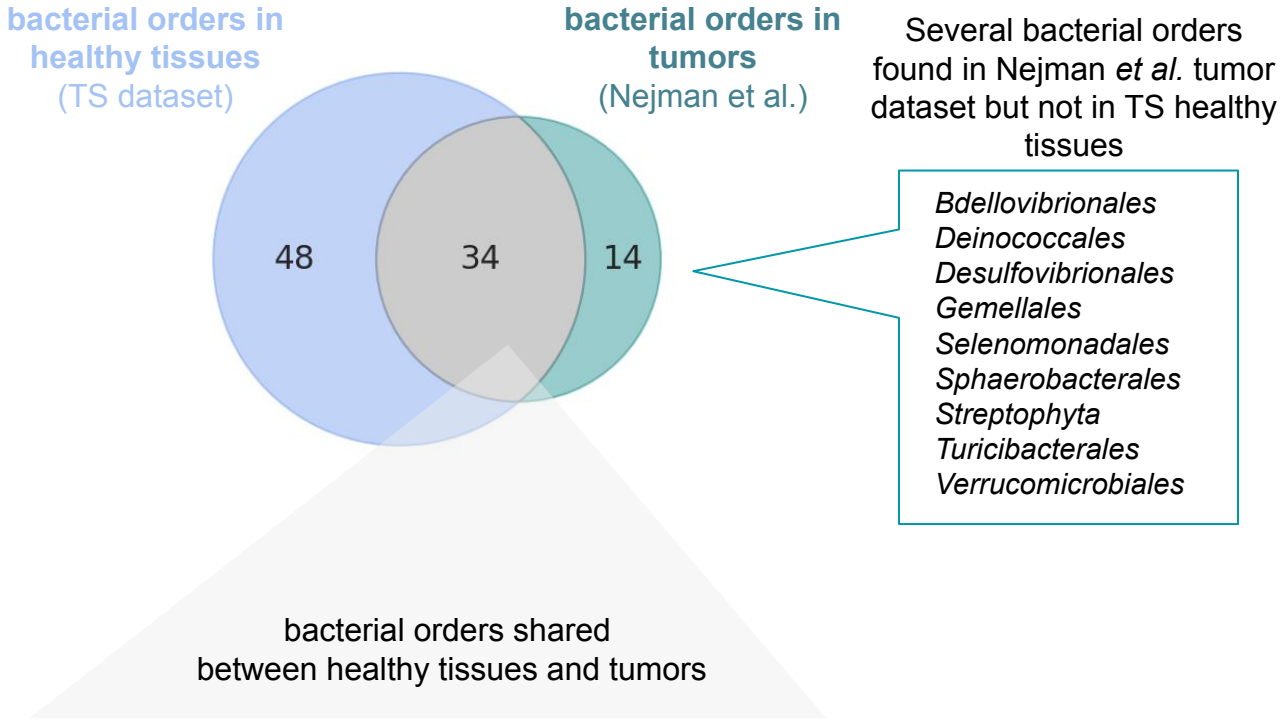

Tissue microbiomes (TS)
External-facing microbiomes (HMP)
Bacterial species
Achromobacter sp.
Acinetobacter baumannii
Acinetobacter haemolyticus
Acinetobacter johnsonii
Acinetobacter junii
Acinetobacter radioresistens
Acinetobacter sp.
Actinomyces graevenitzii
Actinomyces johnsonii
Actinomyces viscosus
Aerococcus urinae
Aeromonas veronii
Aggregatibacter aphrophilus
Aggregatibacter segnis
Anaerococcus prevotii
Anaerococcus sp.
Anaerostipes caccae
Bacillus coagulans
Bacteroides ovatus
Bifidobacterium longum
Campylobacter ureolyticus
Capnocytophaga gingivalis
Citrobacter freundii
Citrobacter koseri
Clostridiales bacterium
Clostridium perfringens
Clostridium sp.
Coprococcus catus
Corynebacterium accolens
Corynebacterium ammoniagenes
Corynebacterium aurimucosum
Corynebacterium jeikeium
Corynebacterium kroppenstedtii
Corynebacterium lipophiloflavum
Corynebacterium matruchotii
Corynebacterium simulans
Corynebacterium sp.
Corynebacterium striatum
Corynebacterium tuberculostearicum
Edwardsiella tarda
Eikenella corrodens
Enterobacter cancerogenus
Enterobacter cloacae
Enterobacter hormaechei
Enterobacter sp.
Enterococcus faecalis
Enterococcus faecium
Erysipelothrix rhusiopathiae
Faecalibacterium prausnitzii
Finegoldia magna
Fusobacterium nucleatum
Fusobacterium sp.
Gemella haemolysans
Gemella sanguinis
Globicatella sp.
Granulicatella adiacens
Haemophilus parainfluenzae
Hafnia alvei
Klebsiella oxytoca
Klebsiella pneumoniae
Kytococcus sp.
Lachnoanaerobaculum sp.
Lactobacillus acidophilus
Lactobacillus gasseri
Lactobacillus iners
Lactobacillus jensenii
Lactobacillus johnsonii
Lautropia mirabilis
Leptotrichia sp.
Leuconostoc mesenteroides
Moraxella sp.
Morganella morganii
Neisseria cinerea
Neisseria elongata
Neisseria meningitidis
Neisseria mucosa
Neisseria sicca
Neisseria sp.
Neisseria subflava
Oribacterium sp.
Parvimonas micra
Pasteurella dagmatis
Peptoniphilus harei
Peptoniphilus sp.
Porphyromonas asaccharolytica
Porphyromonas catoniae
P. gingivalis
Porphyromonas gingivalis
Porphyromonas sp.
Prevotella melaninogenica
Prevotella nigrescens
Propionibacterium acidifaciens
Proteobacteria bacterium
Proteus mirabilis
Providencia stuartii
Pseudomonas aeruginosa
Pseudomonas sp.
Raoultella ornithinolytica
Rothia aeria
R. dentocariosa
Rothia dentocariosa
Rothia mucilaginosa
Salmonella sp.
Segniliparus rugosus
Serratia odorifera
Simonsiella muelleri
Slackia exigua
Staphylococcus aureus
Staphylococcus capitis
Staphylococcus caprae
Staphylococcus haemolyticus
Staphylococcus lugdunensis
Staphylococcus simulans
Staphylococcus sp.
Staphylococcus warneri
Streptococcus australis
Streptococcus cristatus
Streptococcus infantis
Streptococcus intermedius
Streptococcus mitis
S. oralis
Streptococcus oralis
Streptococcus parasanguinis
Streptococcus pseudopneumoniae
S. pyogenes
Streptococcus pyogenes
Streptococcus salivarius
Streptomyces sp.
T. denticola
Treponema denticola
Veillonella dispar
Veillonella parvula
Weissella paramesenteroides
bladder
blood
bone_marrow
eye
fat
heart
large_intestine
liver
lung
lymph_node
muscle
pancreas
salivary_gland
skin
small_intestine
spleen
thymus
trachea
uterus
vasculature
airways
ear
gastrointestinal_tract
nose
oral
skin_hmp
urogenital_tract

SI Figure 5. Bacterial species found in both the Tabula Sapiens (TS) dataset and the Human Microbiome project (HMP). Presence is denoted by the light colored boxes. Boxed in red are species that were referred to in the main text.

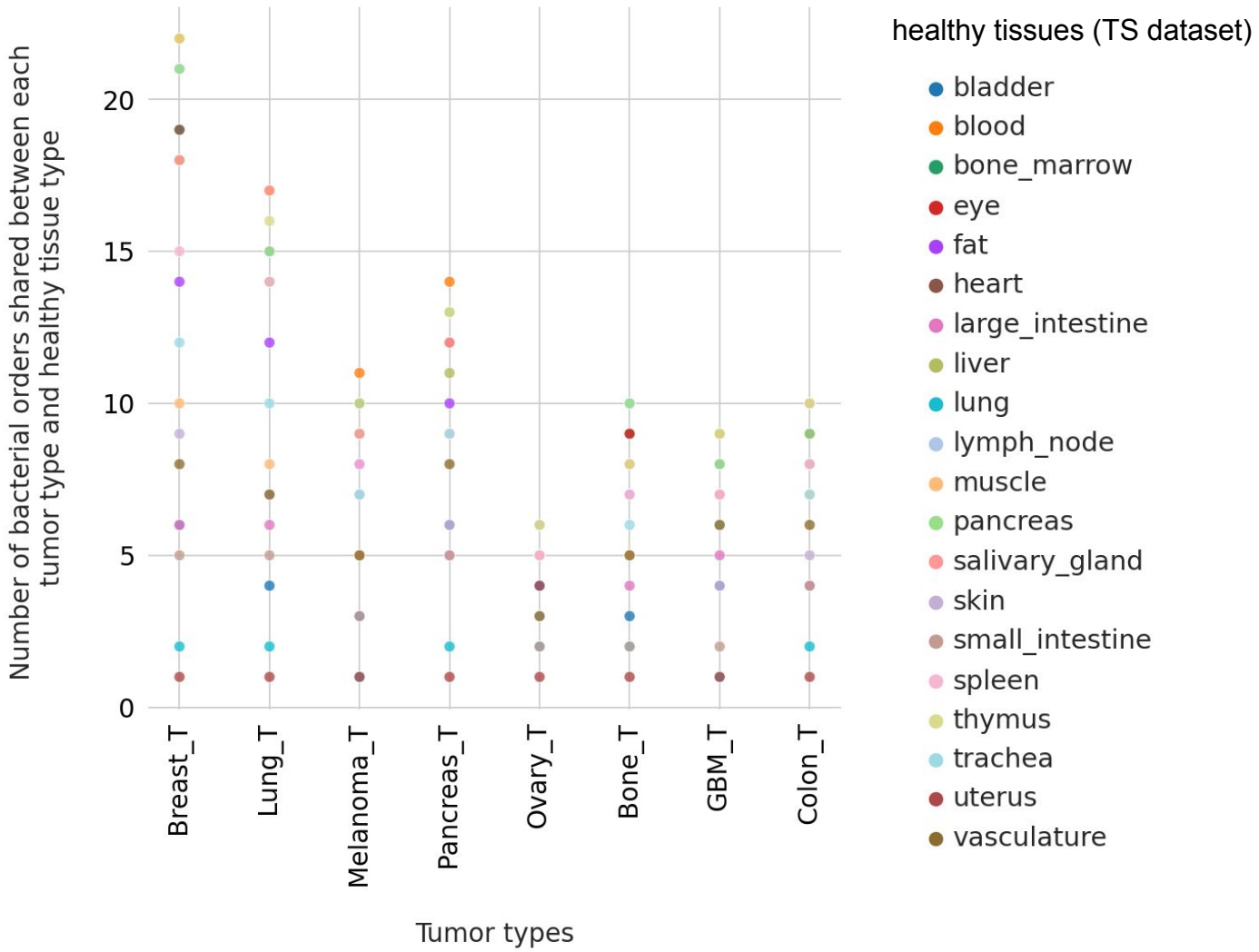

a.
bacterial species in healthy tissues (Tabula Sapiens dataset)
1063
89
189
bacterial species in tumors (Nejman et al.)
b.
Bacterial species found in both healthy tissues and tumors
tissue_donor
c.
Tumor-enriched bacterial species found in healthy tissues
tissue_donor
Staphylococcus aureus
Klebsiella pneumoniae
Fusobacterium nucleatum
Enterobacter cloacae
Staphylococcus cohnii
Log2 (h + 1)

SI Figure 6. Exploring the intersection of bacterial species found in the Tabula Sapiens (TS) dataset and tumor microbiomes from Nejman *et al*. a) Number of species shared between TS and Nejman *et al*. b) abundance of bacterial species shared between the two datasets across different healthy tissues. c) abundance of species that were highly enriched in tumors and also appeared in the TS dataset. Abundance of species is based on Log_2 hits +1 (b,c).

a.


bacterial orders in healthy tissues
(TS dataset)
bacterial orders in tumors
(Nejman et al.)
48
34
14
Several bacterial orders found in Nejman *et al.* tumor dataset but not in TS healthy tissues
Bdellovibrionales
Deinococcales
Desulfovibrionales
Gemellales
Selenomonadales
Sphaerobacterales
Streptophyta
Turicibacterales
Verrucomicrobiales
bacterial orders shared between healthy tissues and tumors

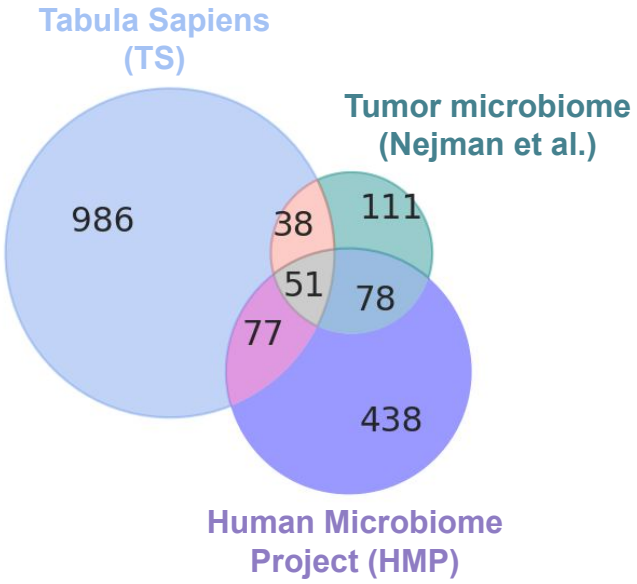
b.


Number of bacterial orders shared between each tumor type and healthy tissue type
0
5
10
15
20
Breast_T
Lung_T
Melanoma_T
Pancreas_T
Ovary_T
Bone_T
GBM_T
Colon_T
Tumor types
healthy tissues (TS dataset)
bladder
blood
bone_marrow
eye
fat
heart
large_intestine
liver
lung
lymph_node
muscle
pancreas
salivary_gland
skin
small_intestine
spleen
thymus
trachea
uterus
vasculature

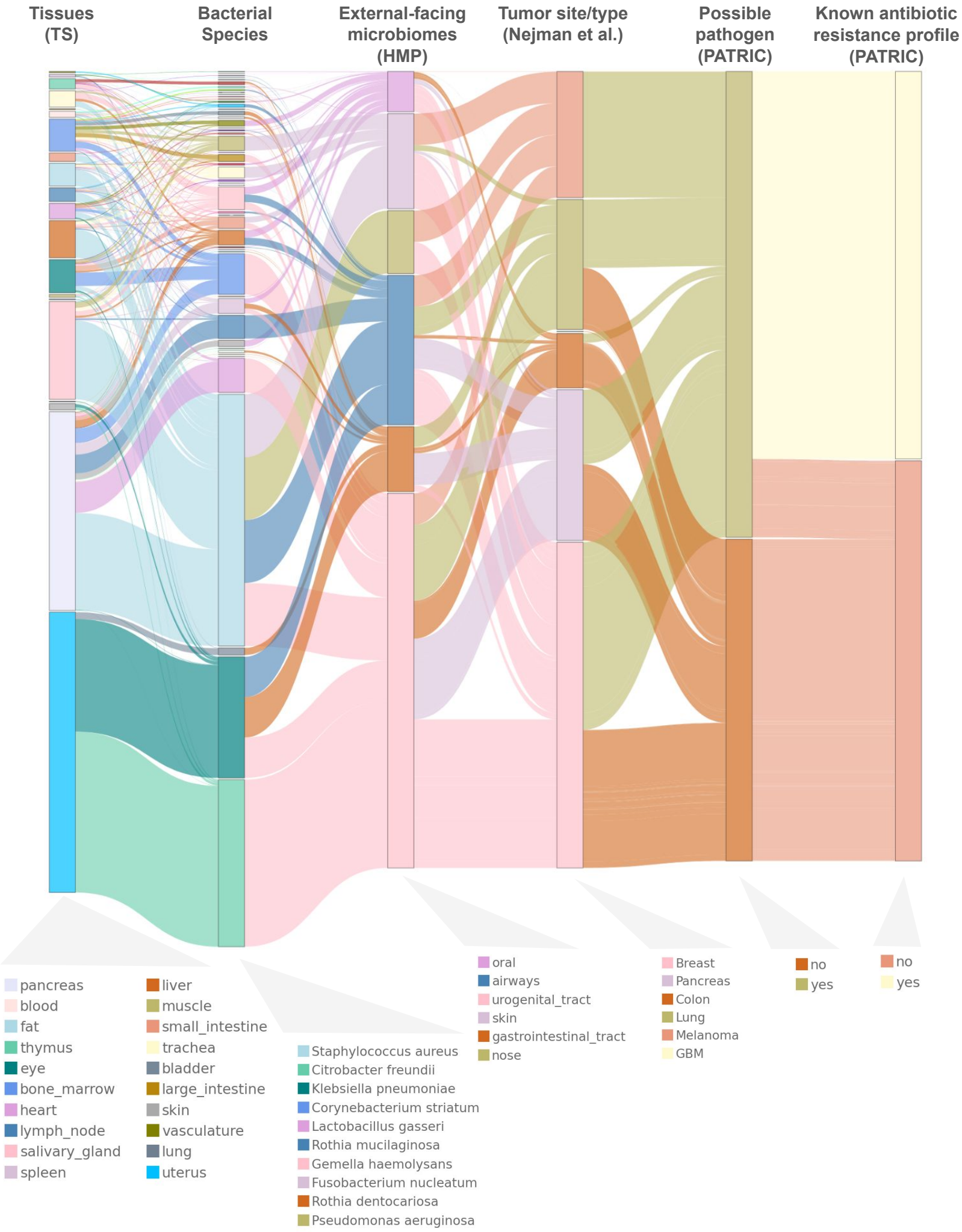
SI Figure 7. Comparison of bacterial orders in healthy tissues (Tabula Sapiens) and tumors (Nejman *et al.*) a) venn-diagram representing the number of shared orders between healthy tissues and tumors. Bacterial orders that are found in tumors but not in healthy tissues are listed in the green box. b) Number of bacterial orders shared between each tumor type and each healthy tissue type is shown.

Healthy tissues
(TS)
Tumors
(Nejman et al.)
Log2 (h + 1)
0
2
4
6
8
10
Alteromonadales
Oceanospirillales
Rhodobacterales
Rhodospirillales
Sphingomonadales
Xanthomonadales
Campylobacterales
Flavobacteriales
Oscillatoriales
Coriobacteriales
Spirochaetales
Rickettsiales
Caulobacterales
Cardiobacteriales
Pirellulales
Vibrionales
Rhodocyclales
Aeromonadales
Sphingobacteriales
Cytophagales
Bifidobacteriales
Erysipelotrichales
Bacteroidales
Clostridiales
Actinomycetales
Fusobacteriales
Neisseriales
Pasteurellales
Burkholderiales
Rhizobiales
Enterobacterales
Bacillales
Lactobacillales
Pseudomonadales
bladder
blood
bone_marrow
eye
fat
heart
large_intestine
liver
lung
lymph_node
muscle
pancreas
salivary_gland
skin
small_intestine
spleen
thymus
trachea
uterus
vasculature
Breast_T
Lung_T
Melanoma_T
Pancreas_T
Ovary_T
Bone_T
GBM_T
Colon_T
Bacterial orders found in both tumors and healthy tissues

SI Figure 8. Abundance of bacterial orders previously found in tumor samples (Nejman *et. al*, 2020) and in the Tabula Sapiens (TS) dataset. Abundance is based on $\log_2$ hits +1 for both heatmaps.

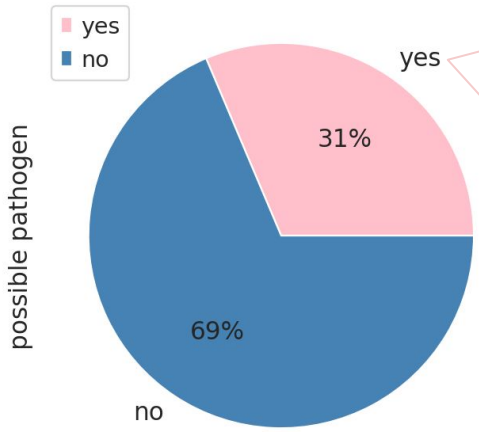

Tabula Sapiens
(TS)
Tumor microbiome
(Nejman et al.)
986
38
111
51
78
77
438
Human Microbiome
Project (HMP)

SI Figure 9. Venn-diagram representing the number of shared bacterial species between the Tabula Sapiens (TS) dataset, the Human Microbiome Project (HMP) and Nejman *et al.* tumor microbiome dataset.

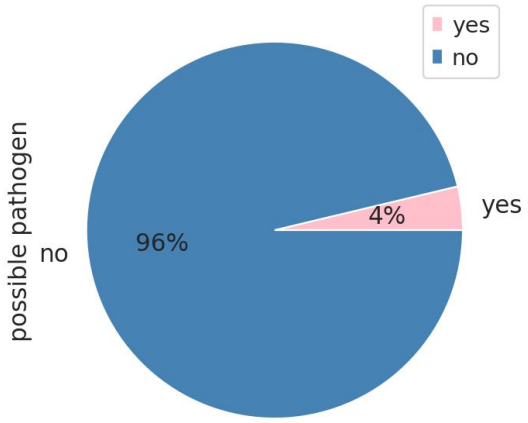

Tissues (TS)
Bacterial Species
External-facing microbiomes (HMP)
Tumor site/type (Nejman et al.)
Possible pathogen (PATRIC)
Known antibiotic resistance profile (PATRIC)
pancreas
blood
fat
thymus
eye
bone_marrow
heart
lymph_node
salivary_gland
spleen
liver
muscle
small_intestine
trachea
bladder
large_intestine
skin
vasculature
lung
uterus
Staphylococcus aureus
Citrobacter freundii
Klebsiella pneumoniae
Corynebacterium striatum
Lactobacillus gasseri
Rothia mucilaginosa
Gemella haemolysans
Fusobacterium nucleatum
Rothia dentocariosa
Pseudomonas aeruginosa
oral
airways
urogenital_tract
skin
gastrointestinal_tract
nose
Breast
Pancreas
Colon
Lung
Melanoma
GBM
no
yes
no
yes

SI Figure 10. Sankey plot of 51 species shared between Tabula Sapiens (TS), Human Microbiome Project (HMP), and Nejman *et al.* datasets. Each line in the plot represents an individual sequence. Only top ten most abundant species (abundance is based on TS dataset) are denoted in the legend. Sequences belonging to species that are not found in the PATRIC database, are labeled as “no” for the last two columns.

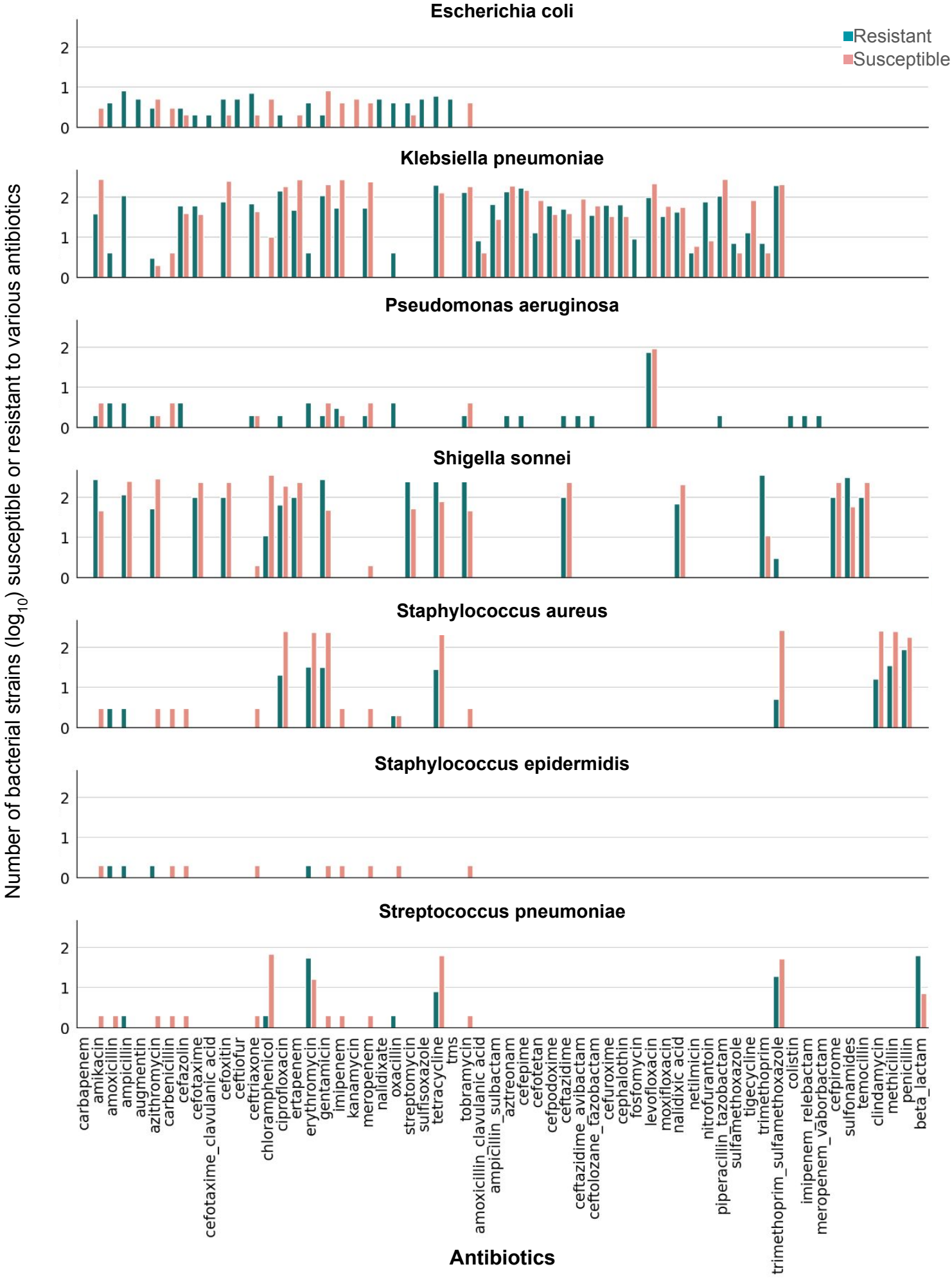
Potentially pathogenic species found in tumors and healthy tissues



Actinomyces viscosus
Corynebacterium aurimucosum
Rothia mucilaginosa
Capnocytophaga gingivalis
Staphylococcus aureus
Staphylococcus haemolyticus
Gemella haemolysans
Enterococcus faecium
Clostridium perfringens
Veillonella parvula
Fusobacterium nucleatum
Eikenella corrodens
Neisseria subflava
Enterobacter cancerogenus
Klebsiella pneumoniae
Pseudomonas aeruginosa

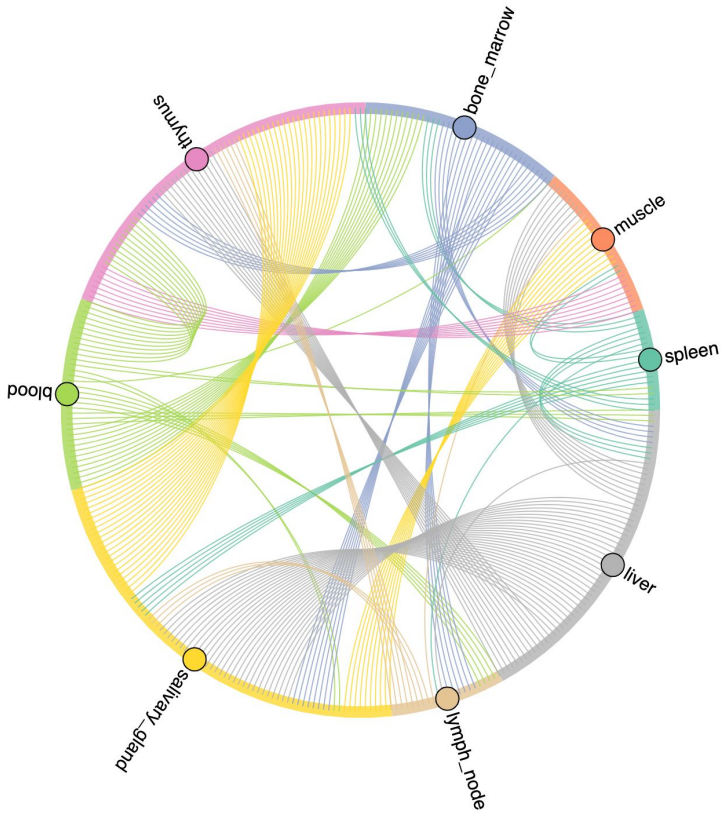


SI Figure 11. Percentage of bacterial species that are potential pathogens based on data from the PATRIC database. a) Percentage of potentially pathogenic species across those species (n=51) found in the Tabula Sapiens, Human Microbiome Project, and Nejman *et al.* datasets. b) Percentage of potentially pathogenic bacterial species across the Tabula Sapiens dataset (n=1152).

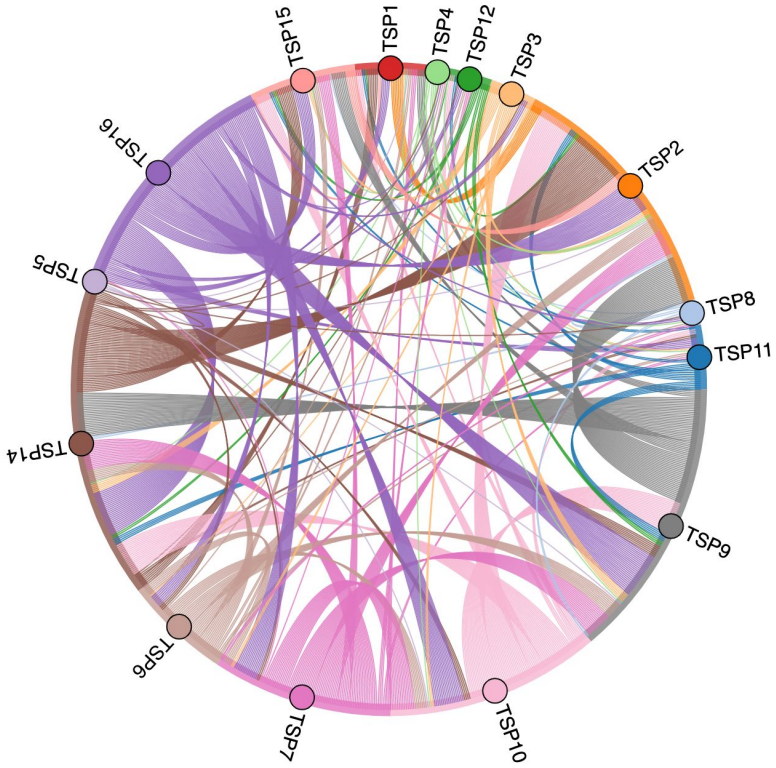

Escherichia coli
Resistant
Susceptible
Klebsiella pneumoniae
Pseudomonas aeruginosa
Shigella sonnei
Staphylococcus aureus
Staphylococcus epidermidis
Streptococcus pneumoniae
Number of bacterial strains (log10) susceptible or resistant to various antibiotics
0
1
2
carbapenem
amikacin
amoxicillin
ampicillin
augmentin
azithromycin
carbenicillin
cefazolin
cefotaxime
cefotaxime_clavulanic acid
cefoxitin
ceftiofur
ceftriaxone
chloramphenicol
ciprofloxacin
ertapenem
erythromycin
gentamicin
imipenem
kanamycin
meropenem
nalidixate
oxacillin
streptomycin
sulfisoxazole
tetracycline
tms
tobramycin
amoxicillin_clavulanic acid
ampicillin_sulbactam
aztreonam
cefepime
cefotetan
cefpodoxime
ceftazidime
ceftazidime_avibactam
ceftolozane_tazobactam
cefuroxime
cephalothin
fosfomycin
levofloxacin
moxifloxacin
nalidixic acid
netilmicin
nitrofurantoin
piperacillin_tazobactam
sulfamethoxazole
tigecycline
trimethoprim
trimethoprim_sulfamethoxazole
colistin
imipenem_relebactam
meropenem_vaborbactam
cefpirome
sulfonamides
temocillin
clindamycin
methicillin
penicillin
beta_lactam
Antibiotics

SI Figure 12. Antimicrobial resistance profile against 60 antibiotics for seven bacterial species (data was sourced from the PATRIC database). Number of bacterial strains (log_{10}) within each species resistant or susceptible to each antibiotic is shown. Refer to SI table 5 for strain-level antibiotic resistance profiles.

a. TSP14 microbiome grouped by tissue

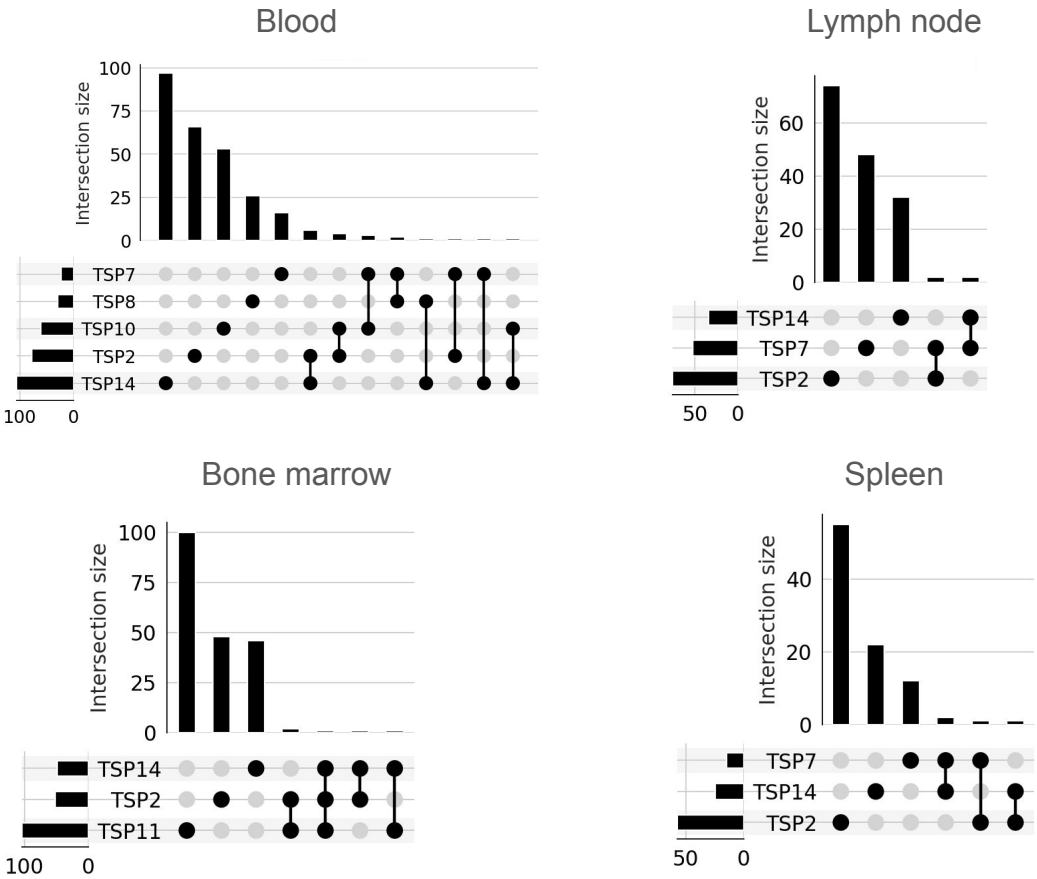


b.

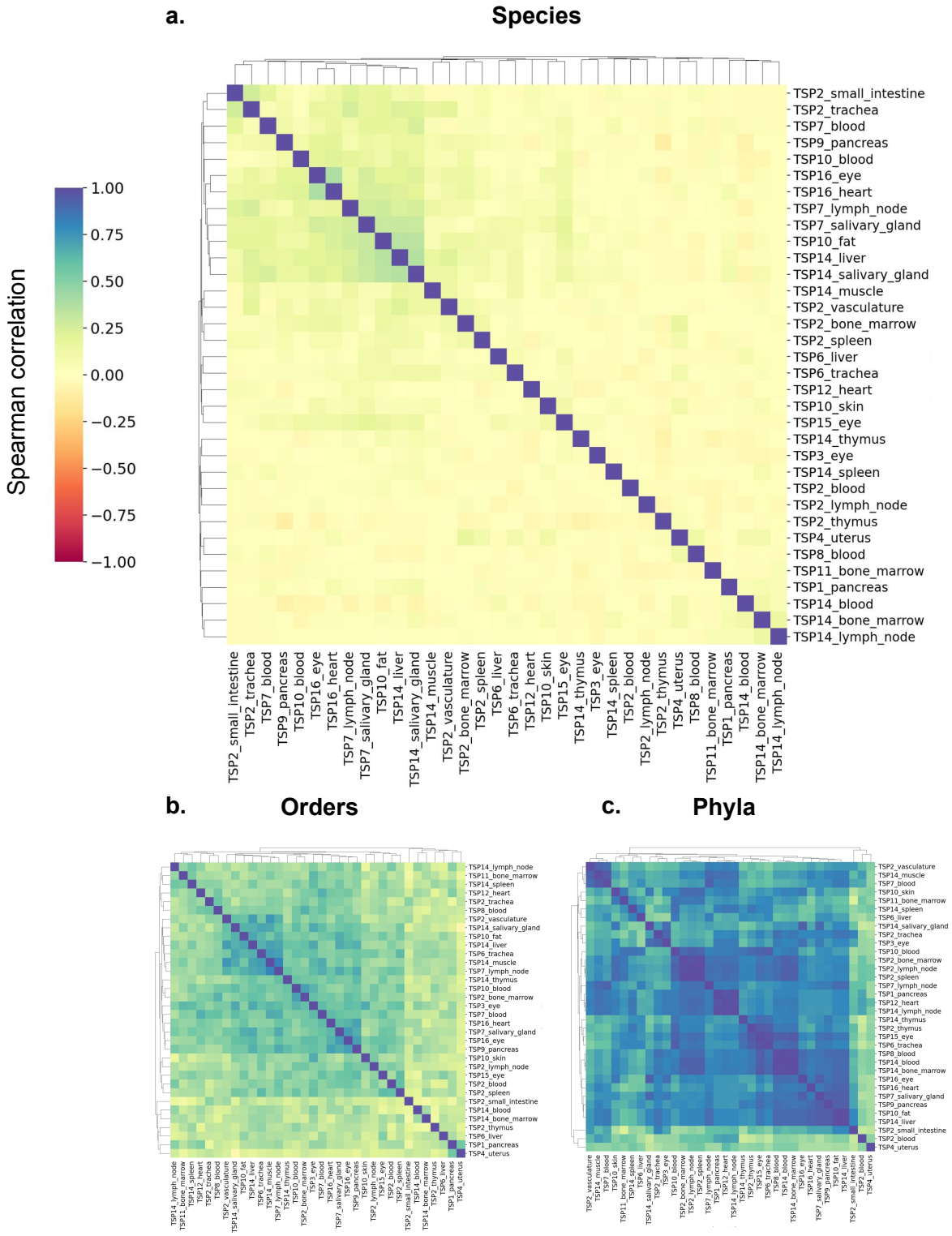


Tissue microbiomes grouped by donor

SI Figure 13. Circos plots of shared species within (a) and between individuals (b). Each link represents a unique species. Species that are not shared are not shown.

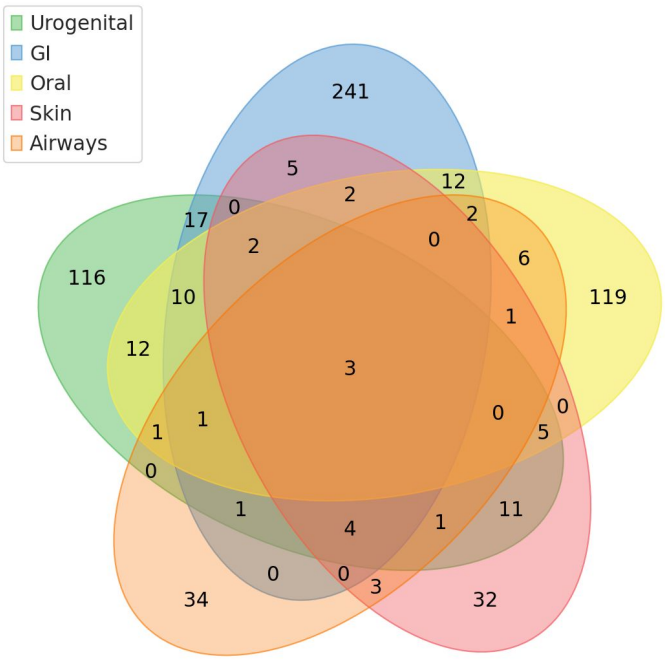

Blood
Intersection size
100
75
50
25
0
TSP7
TSP8
TSP10
TSP2
TSP14
100
0
Lymph node
Intersection size
60
40
20
0
TSP14
TSP7
TSP2
50
0
Bone marrow
Intersection size
100
75
50
25
0
TSP14
TSP2
TSP11
100
0
Spleen
Intersection size
40
20
0
TSP7
TSP14
TSP2
50
0

SI Figure 14. UpSet plots depicting the number of species shared and unique to each sample. UpSet plots drawn for blood, lymph node, bone marrow, and spleen across several donors. Unique sets of species are shown as filled circles. The intersection of two or more sets are shown via connected circles. Bars denote the size of each set or the intersection of sets.


a.
Species
Spearman correlation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
TSP2_small_intestine
TSP2_trachea
TSP7_blood
TSP9_pancreas
TSP10_blood
TSP16_eye
TSP16_heart
TSP7_lymph_node
TSP7_salivary_gland
TSP10_fat
TSP14_liver
TSP14_salivary_gland
TSP14_muscle
TSP2_vasculature
TSP2_bone_marrow
TSP2_spleen
TSP6_liver
TSP6_trachea
TSP12_heart
TSP10_skin
TSP15_eye
TSP14_thymus
TSP3_eye
TSP14_spleen
TSP2_blood
TSP2_lymph_node
TSP2_thymus
TSP4_uterus
TSP8_blood
TSP11_bone_marrow
TSP1_pancreas
TSP14_blood
TSP14_bone_marrow
TSP14_lymph_node
b.
Orders
c.
Phyla

SI Figure 15. Tissue microbiomes across tissues and donors at the a) species, b) order and c) phylum-level of classification. Pairwise Spearman correlation clustermaps across samples grouped by donor and tissue are shown.

Human Microbiome Project (HMP)

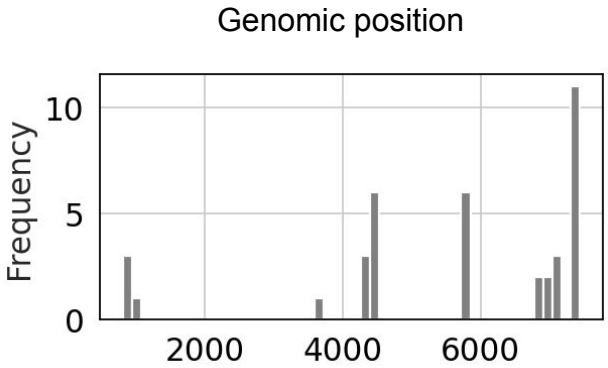


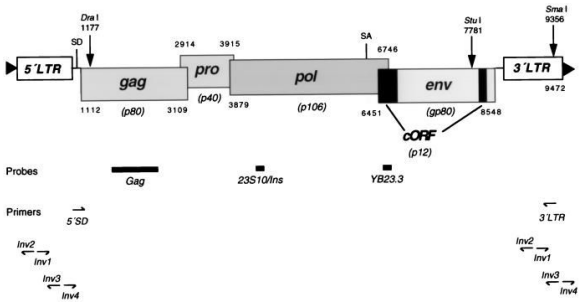
SI Figure 16. Number of species shared across HMP isolation sites (urogenital and GI tracts, oral cavity, skin, and the airways).

a.

donor	tissue	Cell type
TSP10	Fat	endo. c.
		macrophage
	Skin	cd4+, alpha-beta mem. t c.
TSP14	Muscle	tendon c.
	Salivary_Gland	acinar c. of salivary gland
		plasma c.
	Spleen	cd4+, alpha-beta mem. t c.
		mem. b c.
TSP5	Eye	conjunctival epithelial c.
		melanocyte
TSP7	Lymph_Node	b c.
	Salivary_Gland	acinar c. of salivary gland
		mem. b c.
	Spleen	cd8+, alpha-beta mem. t c.
		mem. b c.
		naive cd4+, alpha-beta t c.
		nk c.
		plasma c.
TSP9	Pancreas	t c.

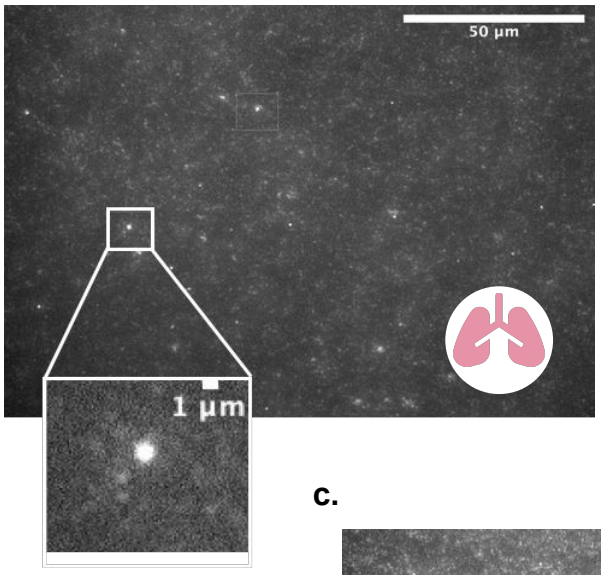
b.

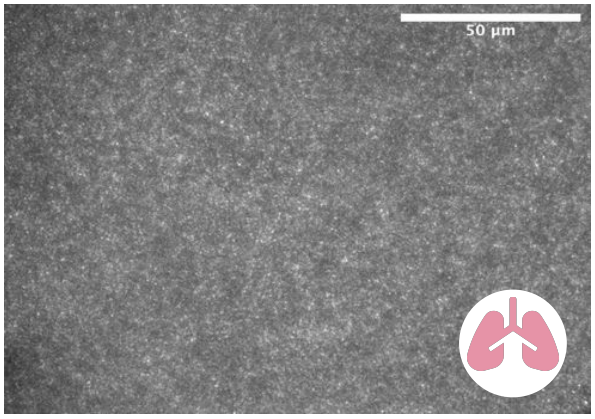


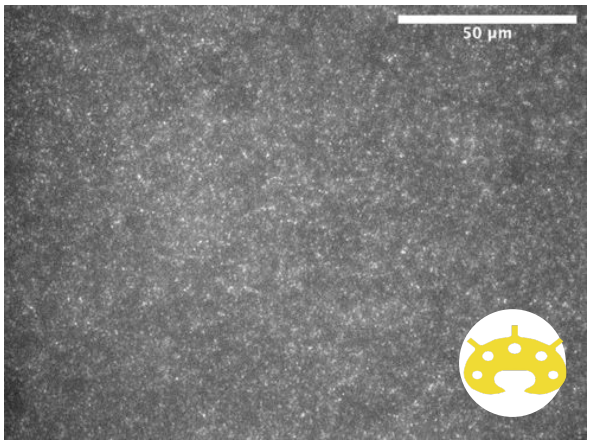


https://pubmed.ncbi.nlm.nih.gov/10516026/

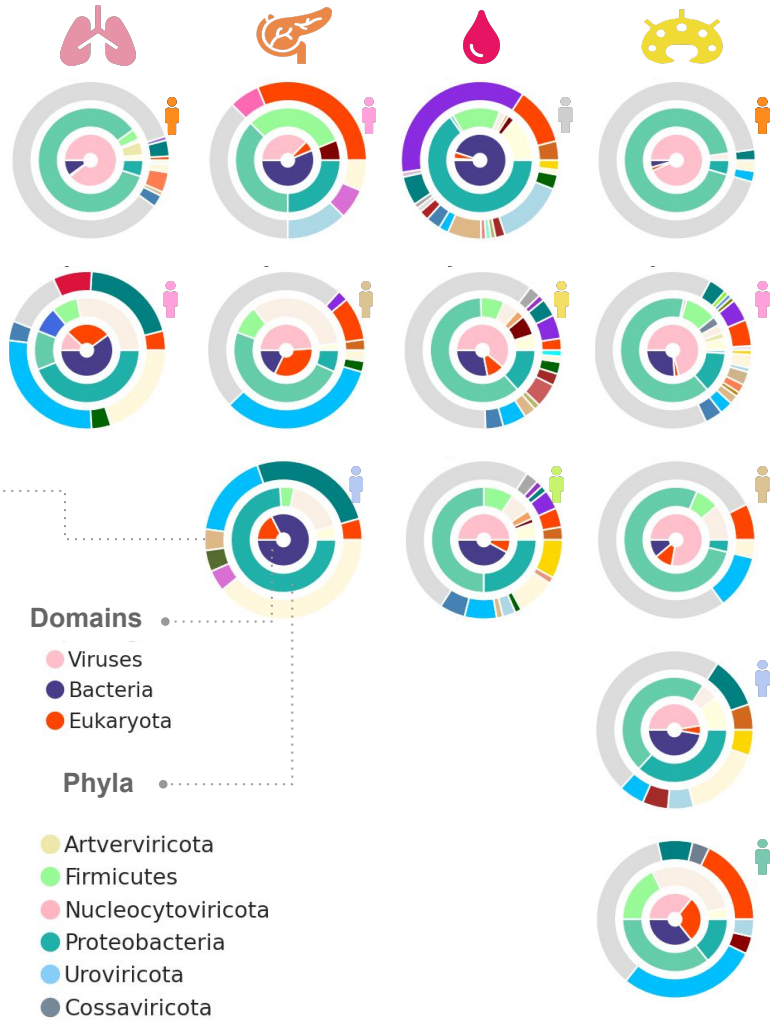
SI Figure 17. Presence of HERV K transcripts across donors, tissues and cell types. a) table summarizing where HERV K sequences were identified. b) the genomic position of identified HERV K hits with respect to the annotated HERV K genome (Tönjes, 1999).


a.
50 µm
1 µm


b.
50 µm


c.
50 µm

SI Figure 18. Fluorescence staining of donor-derived BLPs from the lung (a) and VLPs from the lung and a lymph node (b,c). These images were used to estimate VLP and BLP concentrations in a cubic centimeter of tissue (~10^8 VLPs/cm^3 and ~10^7 BLPs/cm^3).



Orders

- Ortervirales
- Clostridiales
- Algavirales
- Enterobacterales
- Caudovirales
- Piccovirales
- Bacillales
- Nitrosomonadales
- Rhizobiales
- Burkholderiales
- Caulobacterales
- Pseudomonadales
- Nakamurellales
- Propionibacteriales
- Micrococcales
- Lactobacillales
- Actinomycetales
- Sphingomonadales
- Xanthomonadales
- Fusobacteriales
- Micromonosporales
- Flavobacteriales
- Pasteurellales
- Neisseriales
- Veillonellales
- Rhodobacterales
- Geodermatophilales
- Nevskiales
- Coriobacteriales
- Streptosporangiales
- Corynebacteriales
- Malasseziales
- Hypocreales
- Herpesvirales
- Sporidiobolales
- Helotiales
- Pleosporales
- Eurotiales

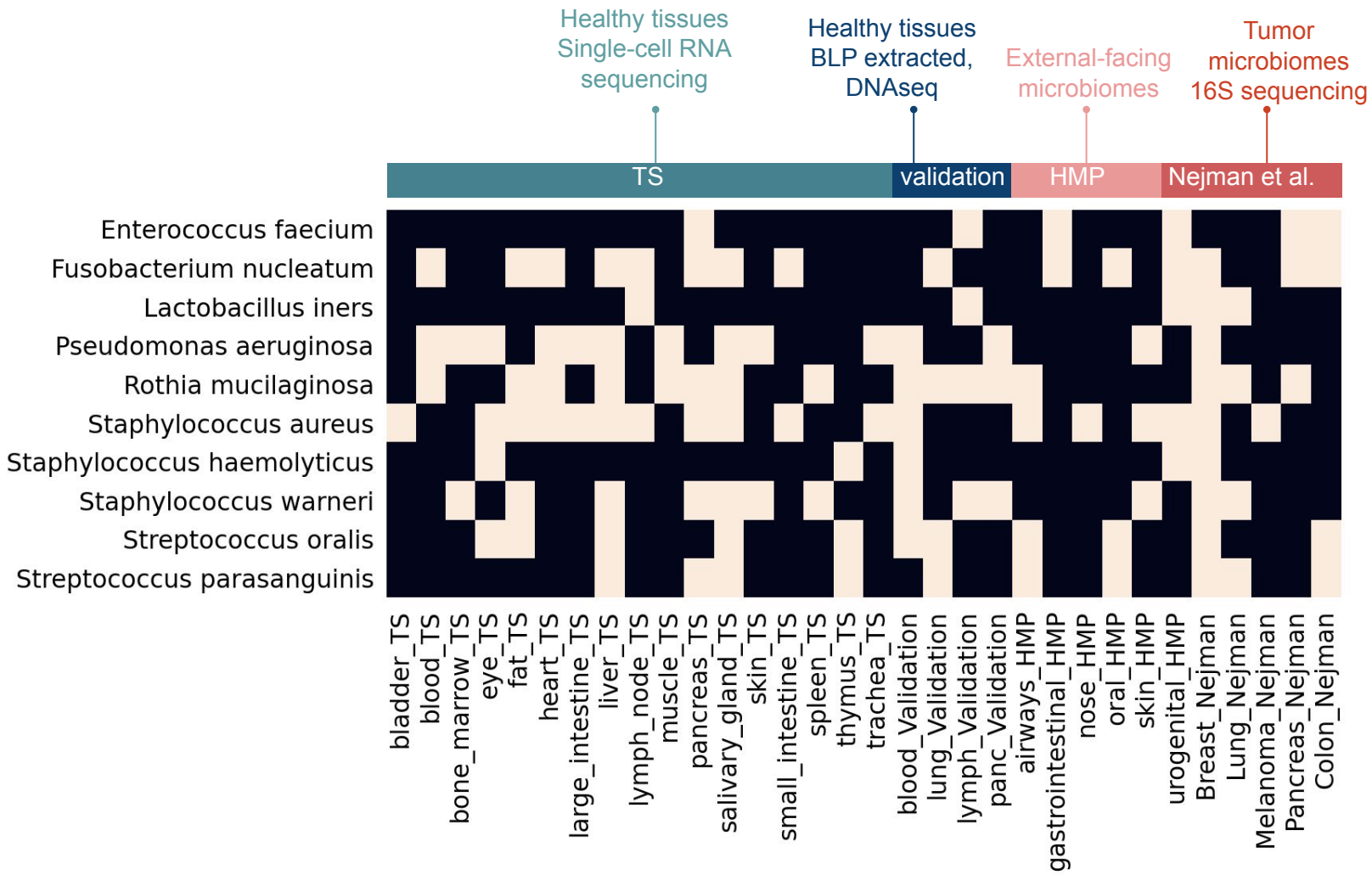
Domains

- Viruses
- Bacteria
- Eukaryota

Phyla

- Artverviricota
- Firmicutes
- Nucleocytoviricota
- Proteobacteria
- Uroviricota
- Cossaviricota
- Actinobacteria
- Fusobacteria
- Bacteroidetes
- Basidiomycota
- Ascomycota
- Peploviricota

SI Figure 19. Main Figure 5c with legend for orders.


Healthy tissues
Single-cell RNA
sequencing
Healthy tissues
BLP extracted,
DNAseq
External-facing
microbiomes
Tumor
microbiomes
16S sequencing
TS
validation
HMP
Nejman et al.
Enterococcus faecium
Fusobacterium nucleatum
Lactobacillus iners
Pseudomonas aeruginosa
Rothia mucilaginosa
Staphylococcus aureus
Staphylococcus haemolyticus
Staphylococcus warneri
Streptococcus oralis
Streptococcus parasanguinis
bladder_TS
blood_TS
bone_marrow_TS
eye_TS
fat_TS
heart_TS
large_intestine_TS
liver_TS
lymph_node_TS
muscle_TS
pancreas_TS
salivary_gland_TS
skin_TS
small_intestine_TS
spleen_TS
thymus_TS
trachea_TS
blood_Validation
lung_Validation
lymph_Validation
panc_Validation
airways_HMP
gastrointestinal_HMP
nose_HMP
oral_HMP
skin_HMP
urogenital_HMP
Breast_Nejman
Lung_Nejman
Melanoma_Nejman
Pancreas_Nejman
Colon_Nejman

TS

SI Figure 20. Species found in all four databases: Tabula Sapiens (TS), Human Microbiome Project (HMP), Nejman *et al.* tumor microbiome, and the validation datasets. Presence (light boxes) and absence (dark boxes) of each species is shown.

SI Table captions

SI Table 1. Taxonomic information, donor, tissue of origin, percent identity and alignment length for ~30,000 microbial sequences found in the Tabula Sapiens dataset.

SI Table 2. Abundance of bacterial species based on ribosomal RNA sequences across different samples (grouped based on donor and tissue). Abundance of species is based on Log_2 hits +1.

SI Table 3. Abundance of bacterial species shared between the two TS dataset and Nejman *et al.* across different healthy tissues. Abundance of species is based on Log_2 hits +1.

SI Table 4. Possible flow paths bacterial species from external-facing microbiomes to internal tissues and tumors. Only flow paths for 51 species shared between the Tabula Sapiens (TS), Human Microbiome Project (HMP) and Nejman et al. datasets are shown.

SI Table 5. Strain-level antibiotic resistance profiles for 1200+ bacterial strains against 60+ antibiotics.

SI Table 6. List of genera found in both bulk DNA sequencing of extracted VLPs and BLPs from 8 donors, that are also identified in single-cell transcriptomic dataset of 15 additional donors.