bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Title: The Tabula Sapiens: a single cell transcriptomic atlas of multiple organs from individual human donors

Authors: The Tabula Sapiens Consortium

One Sentence Summary: We used single cell transcriptomics to create a molecularly defined phenotypic reference of human cell types which spans 24 human tissues and organs.

Abstract: In recent years there has been tremendous progress towards deep molecular characterization of cell types using single cell transcriptome sequencing. Here we report a single cell transcriptomic atlas comprising nearly 500,000 cells from 24 different human tissues and organs. In several instances multiple organs were analyzed from the same donor. Analyzing organs from the same individual controls for genetic background, age, environment, and epigenetic effects, and enables a detailed comparison of cell types that are shared between tissues. This resource provides a rich molecular characterization of more than 400 cell types, their distribution across tissues, and detailed information about tissue specific variation in gene expression. We have used the fact that multiple tissues came from the same donor to study the clonal distribution of T cells between tissues, to understand the tissue specific mutation rate in B cells, and to analyze the cell cycle state and proliferative potential of shared cell types across tissues. Finally, we have also used this data to characterize cell type specific RNA splicing and how such splicing varies across tissues within an individual.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Main Text:

Introduction

Although the genome is often called the blueprint of an organism, it is perhaps more accurate to describe it as a parts list composed of the various genes which may or may not be used in the different cell types of a multicellular organism. Despite the fact that nearly every cell in the body has the same genome, each cell type makes different use of that genome and expresses a subset of all possible genes. Therefore the genome in and of itself does not provide an understanding of the molecular complexity of the various cell types of that organism. This has motivated various efforts to characterize the molecular composition of various cell types within humans and multiple model organisms, both by transcriptional[1] and proteomic[2,3] approaches.

While such efforts are already yielding important insights and a vast amount of data[4–6], one caveat to current approaches is that individual organs are often collected at different locations, from different donors and processed using different protocols, or lack replicate data.[7] Controlled comparisons of cell types between different tissues and organs are especially difficult when donors differ in genetic background, age, environmental exposure, and epigenetic effects. To address this, we have previously developed an approach to analyzing large numbers of organs from the same individual animal using teams of tissue experts who work in coordination with each other,[8] which we used to characterize age-related changes in gene expression in various cell types in the model organism mouse.[9]

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Here we extend such an approach to human organ donors. We have compiled a single cell transcriptomic atlas comprising nearly 500,000 cells from 24 different human tissues and organs. In several instances multiple organs were analyzed from the same donor, thus enabling a detailed within-individual comparison of cell types that are shared between tissues. On a per tissue compartment basis, the data set includes 264,009 immune cells, 102,580 epithelial cells, 32,701 endothelial cells and 81,529 stromal cells.

Data Collection and Cell Type Representation

We collected multiple tissues from individual donors (designated TSP 1-15) and performed coordinated single cell transcriptome analysis on live cells. We collected 17 tissues from one donor, 14 tissues from a second donor, and 5 tissues from two other donors (**Fig. 1**). We also collected smaller numbers of tissues from a further 11 donors, which enabled us to analyze biological replicates for nearly all tissues. The donors comprise a range of ethnicities, are balanced by gender, and have a mean age of 51 years (**Table S1**). Tissues were processed consistently across all donors. Fresh tissues were collected from consented brain-dead transplant donors through an organ procurement organization (OPO) and transported immediately to tissue experts where each tissue was dissociated. As described in the Methods, for many tissues the dissociated cells were purified into compartment-level batches (immune, stromal, epithelial and endothelial) and then recombined into balanced cell suspensions in order to enhanced sensitivity for rare cell types (**Methods**).

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Single cell transcriptome analysis and annotation were performed as outlined in **Fig. S1**. Sequencing was performed using both FACS sorted cells into well plates with smartseq2 amplification as well as 10x microfluidic droplet capture and amplification for each tissue. The raw data was processed to remove low quality cells, projected into a lower-dimensional latent space using scVI, and visualized with UMAP. (**Fig. S2A-E**). Next, the tissue experts used cellxgene[10] to annotate the cells that could be confidently identified by marker gene expression (**Methods**). These annotations were verified through a combination of automated annotation[11] and further manual inspection (**Methods**). A defined Cell Ontology terminology was used to make the annotations consistent across the different tissues, leading to a total of 475 distinct cell types for which we have reference transcriptome profiles (**Table S2**). The full data set can be explored online using the cellxgene tool via a data portal located at tabula-sapiens-portal.ds.czbiohub.org.

Data was collected for bladder, blood, bone marrow, eye, fat, heart, kidney, large intestine, liver, lung, lymph node, mammary, muscle, pancreas, prostate, salivary gland, skin, small intestine, spleen, thymus, tongue, trachea, uterus and vasculature. 59 separate specimens in total were collected, processed, and analyzed, and 481,120 cells passed QC filtering (**Figs. S3 to S7**, **Table S2**). Working with live cells as opposed to isolated nuclei ensured that the dataset includes all mRNA transcripts within the cell, including transcripts that have been processed by the cell's splicing machinery, thereby enabling insight into the broad and ubiquitous phenomenon of alternative splice variation.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

For several of the tissues we also performed literature searches and collected tables of prior knowledge of cell type identity and abundance within those tissues (**Table S3**). We compared those literature values with our experimentally observed frequencies for three well annotated tissues: lung, muscle and bladder (**Fig. S8**). There is surprisingly good correspondence in the frequencies, especially considering that the single cell data was obtained on tissues that were dissociated and that various compartments were enriched.

To further characterize the relationship between transcriptome data and conventional histologic analysis of tissue, a team of trained pathologists analyzed H&E stained sections prepared from 9 tissues from donor TSP 2 and 13 tissues from donor TSP14 (**Data Portal**). Cells were identified by morphology and classified broadly into epithelial, endothelial, immune and stromal compartments, as well as rarely detected peripheral nervous system (PNS) cell types. In some cases, finer cell type classification was also performed. An example of such cellular and compartmental identification is illustrated in the case of the distal small intestine (**Fig. 2A**). These classifications were used to estimate the relative abundances of cell types across four compartments, as well as to the uncertainties in these abundances due to spatial heterogeneity of each tissue type. (**Fig. 2B)** We compared the histologically determined abundances with those obtained by single cell sequencing (**Fig. 2C**). Although as expected there can be substantial variation between the abundances determined by these methods, we do in aggregate observe broad concordance over a large range of tissues and relative abundances. This approach enables an estimate of true cell type proportions for organs where the

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

compartments were purified, and more generally in every organ since not every cell type survives dissociation with equal efficiency.[12] The histology images of the tissue are available as part of the Tabula Sapiens Data Portal.

Immune Cells: Variation in Gene Expression Across Tissues and a Shared Lineage History

The Tabula Sapiens can be used to study subtle differences in the gene expression programs and lineage histories of cell types that are shared across tissues. Importantly, these analyses were performed after correcting counts for potential ambient mRNA contamination and dissociation artifacts (**Methods**), which would otherwise result in the detection of differentially expressed genes (DEGs) that are specific to a tissue rather than to the cell type of interest. We first examined immune cells, which are born in one niche, circulate through the body, and home to other niches where they stay for time scales of minutes to years. We first compared different immune cell subsets across different tissues to understand the molecular details of this trafficking within an individual. We identified tissue-specific gene expression features for most immune cell ontology classes via classical DEG analysis. Here we focus on the signatures of tissue similarity and differences in the 36,475 macrophages distributed amongst 20 tissues, as tissue-resident macrophages are known to carry out specialized functions in different tissues and under different conditions. These shared and orthogonal signatures are summarized in a correlation map (**Fig. S10A**). For example, macrophages in the spleen were quite different from most other macrophages, and this difference was driven largely by higher expression of CD5L (**Fig. S10B**). We also observed a shared

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

signature of elevated EREG expression in solid tissues such as the skin, uterus and mammary compared with EREG expression in circulatory tissues (**Fig. S10B**). Macrophages are thought to secrete abnormal levels of EREG during cancer-progression and facilitate the tumor micro-environment,[13] but secretion of such factors also has an important role in homeostatic maintenance of tissues.[14] We also observed a weak correlation of an antimicrobial phenotype of macrophage in the lung and lymph node characterized by CHIT1 expression (**Fig. S10B**). Interestingly, macrophages in the lymph nodes co-expressed CHIT1 and CTSK, while CTSK was largely not expressed in the lung (**Fig. S10B**). Like EREG, CTSK is thought to have roles in cancer metastasis as well as normal tissue regulation. Together, these data provide insight into tissue-specific specializations and functional differences of macrophages.

To characterize the lineage relationships between T cells found in various organs we performed computational assembly of the T cell receptor sequence from T cells sequenced via Smartseq2 from donor TSP2. We discovered that multiple T cell lineages were distributed across various tissues in the body, and mapped their relationships **(Fig. 3A)**. Large clones were often found to reside in multiple organs (**Fig. 3A, S10C**). We found that several clones of Mucosal Associated Invariant T cells where shared across donors; these cells were identified by their characteristic expression of *TRAV1-2* and are thought to be innate-like effector cells.[15]

Lineage information can also be used to measure the level of tissue-specific somatic hyper-mutation in B cells. We computationally assembled the B Cell Receptor (BCR)

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

gene from Smartseq2 data from donor TSP2 and then inferred the germline ancestor of each cell. In the spleen and lymph nodes we observed a bimodal distribution of hypermutation, consistent with the coexistence of plasma cells and recently birthed B cells **(Fig. 3B)**. Solid tissues have an order of magnitude more mutations per nucleotide (mean=0.076, s.d.=0.026) compared to the blood (0.0069), suggesting that the immune infiltrates of solid tissues are dominated by mature B cells (**Fig. S10D**).

B cells also undergo class-switch recombination which diversifies the humoral immune response by using constant region genes with distinct roles in immunity. We classified every B cell in the dataset as IgA, IgG, or IgM expressing and then calculated the relative amounts of each cellular isotype in each tissue. Secretory IgA is known to interact with pathogens and commensals at the mucosae, IgG is often involved in direct neutralization of pathogens, and IgM is typically expressed in naive B cells or secreted in first response to pathogens. Consistent with these functions, our analysis revealed opposing gradients of prevalence of IgA and IgM expressing B cells across the tissues with blood having the lowest relative abundance of IgA producing cells and the large intestine having the highest relative abundance, and the converse for IgM expressing B cells (**Fig. 3B**).

Endothelial Cells Subtypes with Tissue-Specific Gene Expression Programs

As another example application of using the Tabula Sapiens to analyze shared cell types across organs, we focused on endothelial cells (ECs). These cells line the surface of blood vessels and together form a conduit allowing for inter-tissue

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

communication, oxygen, nutrient and waste exchange, and tissue-level homeostasis. While ECs are widely categorized as a single cell type, they exhibit vast differences in morphology, structure, immunomodulatory and metabolic phenotypes depending on their tissue of origin. Here, we discovered that tissue-specificity is also reflected in their transcriptomes, as ECs mainly cluster by tissue-of-origin. UMAP analysis (**Fig. S11A,**) revealed that lung, heart, uterus, liver, pancreas, fat and muscle ECs exhibited the most distinct transcriptional signatures, reflecting their highly specialized roles. These distributions were conserved across donors (**Fig. S11B**). Interestingly, ECs from the thymus, vasculature, prostate, and eye were similarly distributed across several clusters, suggesting not only similarity in transcriptional profiles but in their sources of heterogeneity. Differential gene expression analysis between ECs of these 16 tissues revealed several canonical and previously undescribed tissue-specific vascular markers (**Fig. 3C**). We recapitulated known tissue-specific vascular markers such as LCN1 (tear lipocalin) in the eye, ABCG2 (transporter at the blood-testis barrier) in the prostate, and OIT3 (oncoprotein induced transcript 3) in the liver. Potential novel markers include KRT14 (keratin 14) in the tongue, FAM13C (family with sequence similariy 13, member C) in the pancreas, CYTL1 (cytoline-like 1) in the bladder, DSG2 (a component of intercellular desmosome junctions) in the fat, F2RL3 (a coagulation factor) in the skin, SLC14A1 (solute carrier family 14 member 1) in the heart, and HILPDA (a hypoxia-inducible protein that stimulates cytokine production) in the uterus. Vascular-bed specific genes could provide further insight into tissue-specific homeostatic mechanisms, as well as allow for EC tissue-specificity to be deconvolved in experiments like flow cytometry.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Notably, lung ECs formed two distinct populations, which is in line with the aerocyte (aCap- EDNRB+) and general capillary (gCap - PLVAP+) cells recently described in the mouse and human lung[16] (**Fig. S11C,D**). The transcriptional profile of gCaps were also more similar to ECs from other tissues, indicative of their general vascular functions in contrast to the more specialized aCap populations. Lastly, we detected two distinct populations of ECs in the muscle, including a MSX1+ population with strong angiogenic and endothelial cell proliferation signatures, and a CYP1B1+ population enriched in metabolic genes, suggesting the presence of functional specialization in muscle vasculature (**Fig. S11E,F**).

Alternative Splice Variants are Cell Type Specific

The Tabula Sapiens can also be used to understand cell type specific usage of alternative splicing. The GRCH38 RefSeq genome annotation contains 37,344 genes with multiple annotated exons, 21,923 of which have multiple annotated transcripts, totaling 169,061 variants.[17] Yet the function of alternative splicing and the extent to which regulation is cell type specific remains largely unexplored. We used SICILIAN, a statistical method that removes false positive spliced alignments due to technical artifacts to identify splice junctions in the Tabula Sapiens corpus. Among other statistical filters, SICILIAN requires each called junction to have at least two supporting reads.[18]

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SICILIAN detected a total of 955,785 junctions (**Fig. S12A-C**). Of these, 217,855 were previously annotated, and thus our data provides independent validation of 61% of the 358,924 total junctions catalogued in the RefSeq database. Although annotated junctions made up only 22.8% of the unique junctions, they represent 93% of total reads, indicating that previously annotated junctions tend to be expressed at higher levels than novel junctions. We additionally found 34,624 novel junctions between previously annotated 3' and 5' splice sites (3.6%). We also identified 119,276 junctions between a previously annotated site and a novel site in the gene (12.4%). This leaves 584,030 putative junctions for which both splice sites were previously unannotated, i.e. about 61% of the total detected junctions. Most of these have at least one end in a known gene (94.7%), while the remainder represent potential new splice variants from totally unannotated regions (5.3%). Plasma cells had the highest proportion of their reads coming from junctions with neither end annotated (4.5% and 8.3% respectively in TSP1 and TSP2), while erythrocytes had the lowest percent (0.21% and 0.24% respectively). While many of these unannotated junctions will require independent confirmation to have full confidence in their existence, these results suggest that a substantial amount of splicing may have been missed in previous studies. Future work will be needed to distinguish which result from stochastic versus reproducible and tightly regulated splicing programs.[19,20]

There were hundreds of highly cell-type specific splice expression patterns, and they can be explored in the cellxgene browser using a statistical approach for detecting cell-type-specific splicing called the SpliZ.[21] Here we focus on two examples of cell type

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

specific usage of two well studied genes: MYL6 and CD47. Both genes are ubiquitously expressed yet are highly regulated in single cells at the level of splicing (**Fig. 4**).

MYL6 is an "essential light chain" (ELC) for myosin (as opposed to Ca2+-sensitive "regulatory light chains" (RLCs) such as calmodulin), and is highly expressed in all tissues and compartments. Yet, splicing of MYL6, in particular involving the inclusion/exclusion of exon 6 (Fig. 4A varies in a cell-type and compartment-specific manner (**Fig. 4A,B**). The two isoforms differ by 5 amino acids in the C-terminal helix, which is in close contact with the myosin lever arm; some studies suggest that the -exon6 isoform confers on myosin a faster shortening velocity.[22] While the -exon6 isoform has previously been mainly described in phasic smooth muscle, [23] the more comprehensive nature of the Tabula Sapiens atlas shows that it can also be the predominant isoform in non-smooth-muscle cell types. Our analysis establishes pervasive regulation of MYL6 splicing in many cell types, such as endothelial and immune cells. Further, we demonstrate previously unknown compartment-specific expression patterns of the two MYL6 isoforms that are reproduced in multiple individuals from the Tabula Sapiens dataset (**Fig. 4A,B**) and using both 10X and Smart-Seq2 sequencing technologies.

CD47 is a multi-spanning membrane protein involved in many cellular processes, including angiogenesis, cell migration, and as a "don't eat me" signal to macrophages. Targeting the latter function has been promising for treating some myeloid malignancies. [24] CD47 has complex splicing patterns that include alternative inclusion

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of at least 4 different exons immediately adjacent to the signaling domain ending at the 3' splice splice site at exon 11 (**Fig. S12D**). Differential use of exons 7-10 (Fig. 4B, S11C) compose a variably long cytoplasmic tail.[25] Immune cells- but also stromal and endothelial cells - have a distinct, consistent splicing pattern that dominantly excludes two proximal exons and splicing directly to exon 8. In contrast to other compartments, epithelial cells exhibit a starkly different splicing pattern that increases the length of the cytoplasmic tail by splicing more commonly to exon 9 and exon 10 (**Fig. 4C,D**). Characterization of the splicing programs of CD47 in single cells may have important implications for understanding the differential signaling activities of CD47 and for understanding therapeutic manipulation of CD47 function.

Cell State Dynamics Can Be Inferred From A Single Time Point

Although the Tabula Sapiens was created from a single moment in time for each donor, it is possible to infer various forms of dynamic information about the cells from the data. For example, one of the most important transient changes of internal cell state is cell division. We computed a cycling index for each cell type across all organs to identify actively proliferating versus quiescent or post-mitotic cell states. This index was derived based on the log ratio of the number of cycling to non-cycling cells for each cell type, determined by high confidence cell cycle markers for for G1-M phases (G1/S markers: CEP57, CDCA7L; S markers: ABHD10, CCDC14, CDKN2AIP, NT5DC1, SVIP, PTAR1; G2 makers: ANKRD36C, YEATS4, DCTPP1; G2/M markers: SMC4, TMPO, LMNB1, HINT3; M markers: HMG20B, HMGB3, HPS4) to indicate cycling and G0 phase markers (CDKN1A, CDKN1B, CDKN1C) for non-cycling (see **Methods**).

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In validation of this approach, and across all donors, we observed that rapidly dividing progenitor cells had among the highest cycling indices, while cell types mostly from the endothelial and stromal compartments, which are known to be largely quiescent, had low cycling indices (**Fig. 5A**). In the intestinal tissue, the transient amplifying cells and the crypt stem cells which divide rapidly in the intestinal crypts to give rise to terminally differentiated cell types of the villi,[26] had the highest ranking cycling indices whereas terminally differentiated cell types such as the goblet cells had the lowest ranks (**Fig. S13A**). To complement the computational analysis of cell cycling, we performed immunostaining of intestinal tissue for MKI67 protein (commonly referred to as Ki-67) and observed that transient amplifying cells abundantly express this proliferation marker (**Fig. S13B**), supporting our finding that this marker is differentially expressed in the G2/M cluster (**Fig. S13C**).

We observed several interesting tissue-specific differences in cell cycling. To illustrate one example, UMAP clustering of macrophages showed tissue-specific clustering of this cell type, and that blood, bone marrow, and lung macrophages have the highest cycling indices compared to macrophages found in the bladder, skin, and muscle (**Fig. S13D-G**). Consistent with this finding, the expression values of CDK-inhibitors (in particular the gene CDKN1A), which block the cell cycle, have the least overall expression in macrophages from tissues with high cycling indices (**Fig. S13F**).

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

As a further example of how the Tabula Sapiens can be used to reveal cell state dynamics, we used RNA velocity[27] to study trans-differentiation of bladder mesenchymal cells to myofibroblasts (**Fig. 5B**). This process is important for tissue remodeling and healing, and if left unchecked can result in fibrosis. Myofibroblasts produce different components of the ECM such as collagen and fibronectin. Latent time analysis, which provides an estimate of each cell's internal clock using RNA velocity trajectories,[28] correctly identified the direction of differentiation without requiring specification of root cells (**Fig. 5C**). Similar trajectories were found across multiple donors. Finally, the ordering of cells as a function of latent time shows a clear clustering of the mesenchymal and myfibroblast gene expression programs for the most dynamically expressed genes (**Fig. 5C)**. Among these genes, ACTN1 (Alpha Actinin 1) - a key an actin crosslinking protein that stabilizes cytoskeleton-membrane interactions[29] - increases across the mesenchymal to myofibroblast trans-differentiation trajectory (**Fig. S13H**). Another gene with a similar trajectory is MYLK (myosin light-chain kinase),[30] which is also expected to rise as myofibroblasts attain more muscle-like properties. Finally, a random sampling of the most dynamic genes shared across TSP1 and TSP2 demonstrated that they share concordant trajectories and revealed some of the core genes in the transcriptional program underlying this trans-differentiation event within the bladder (**Fig. S13I**).

Unexpected Spatial Variation in the Microbiome

Imbalances in the interactions between the gut microbiota and the host immune system impact are linked with many region-specific intestinal,[31,32] and stool is not necessarily

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

representative of the spatially distinct microbial[33] and immune[34] niches throughout the intestinal tract. Despite this importance, the spatial heterogeneity of the microbiome remains understudied and largely unknown. The Tabula Sapiens provided an opportunity to densely and directly sample the human microbiome throughout the gastrointestinal tract. The intestines from donor TSP 2 were sectioned into five regions: the duodenum, jejunum, ileum, and ascending and sigmoid colon (**Fig. 6A**). Each section was transected, and three to five samples of ~1 g of digesta were collected from each location using an inoculating loop, excepting the ileum for which three samples were collected. Due to the nature of sample collection, the digesta was largely from the mucosal region adjacent to the epithelium. Samples were spaced approximately 3 inches along the longitudinal gut axis, and some microbiome samples were collected close to the regions of epithelial tissue collection.

DNA was extracted from all 23 samples and the 16S rRNA gene was amplified and sequenced to determine microbiome composition. Across all samples, there was a high (>30%) relative abundance of Proteobacteria, particularly Enterobacteriaceae (**Fig. 6B**), even in the colon; Enterobacteriaceae are rarely found at such high abundance in stool, hence this high relative abundance may be due to the postmortem state of the donor. Samples within the sigmoid and ascending colon were relatively similar to each other, whereas samples from the duodenum, jejunum, and ileum were highly distinct (**Fig. 6B**). These data reveal that the microbiota is highly heterogeneous along the intestines, even at a 3-inch length scale.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In the small intestine, species richness (number of observed species) was also variable and was negatively correlated with the relative abundance of Enterobacteriaceae (**Fig. 6C**). Shannon diversity largely mimicked the number of observed species (**Fig. 6D**). Comparison of species from adjacent regions in the small intestine showed that a large fraction of species was unique to each region (**Fig. 6E**), reflecting the patchiness of the small intestine. By contrast, a much smaller fraction of the species were unique in comparing the ascending and sigmoid colon. The average number of unique ASVs (a proxy for species) was 10.7±4.2 across samples from the duodenum, jejunum, and ileum, as compared with 1.56±0.79 across samples from the ascending and sigmoid colon regions. Moreover, considering samples from within each coarse-grained region, the duodenum, jejunum, and ileum exhibited 17.0±5.8, 19.4±6.8, and 20±5 unique ASVs respectively, reflecting significant heterogeneity at individual sampling sites, compared with 3.0±1.6 and 3.6±1.5 for the ascending and sigmoid colon, respectively.

Taken together, the ability to densely sample the microbiota throughout the gastrointestinal tract reveals previously unrecognized heterogeneity along with region specificity, highlighting the need for extensive characterization of the spatial variation in microbiota composition within and across humans.

Conclusion

The Tabula Sapiens provides an integrated molecular reference of human cell types from 24 different organs across 14 donors. It enables cross-tissue comparisons of a variety of cell types and demonstrates that many cell types, although broadly shared

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

across tissues, nonetheless have well defined tissue-specific gene expression and splicing profiles that enable tissue-specific subtyping. Our analysis has further enabled understanding of how specific immune cell clones are shared between some tissues and how hypermutation rates amongst B cells are tissue dependent. We have observed quite different behavior of cell types shared across tissues, for example that endothelial cells cluster primarily by organ and that certain immune cells have tissue-dependent differences in gene expression. Finally, this atlas provides a basis for discovering cell-type specific RNA splicing, and important but largely unexplored phenomenon. Although this work is the most comprehensive human cell atlas constructed to date, it represents the first draft of a broadly useful reference to understand and explore human biology deeply at cellular resolution. We expect that, similar to the human genome project, over time the release of updated versions of the Tabula Sapiens will incorporate data from additional donors and include further refinements in the cell type annotations.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References:

1. Regev, A., Teichmann, S. A., Lander, E. S., et al., The human cell atlas. *Elife*, 2017, **6**.
2. Uhlén, M., Fagerberg, L., Hallström, B. M., et al., Tissue-based map of the human proteome. *Science (80-.).*, 2015, **347**.
3. Thul, P. J., Åkesson, L., Wiking, M., et al., A subcellular map of the human proteome. *Science (80-.).*, 2017, **356**.
4. Pisco, A. O., Tojo, B., and McGeever, A., Single-Cell Analysis for Whole-Organism Datasets. *Annu. Rev. Biomed. Data Sci.*, 2021, **4**.
5. Aldridge, S. and Teichmann, S. A., Single cell transcriptomics comes of age. *Nat. Commun.*, Nature Research, 2020, December 1, pp. 1–4.
6. Wu, A. R., Wang, J., Streets, A. M., and Huang, Y., Single-Cell Transcriptional Analysis. *Annu. Rev. Anal. Chem.*, 2017, **10**, 439–462.
7. He, S., Wang, L.-H., Liu, Y., et al., Single-cell transcriptome profiling of an adult human cell atlas of 15 major organs. *Genome Biol. 2020 211*, 2020, **21**, 1–34.
8. Tabula Muris Consortium, Overall coordination, Logistical coordination, et al., Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, 2018, **562**, 367–372.
9. The Tabula Muris Consortium, A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature*, 2020, **583**, 590–595.
10. Megill, C., Martin, B., Weaver, C., et al., cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices. *bioRxiv*, 2021,

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2021.04.05.438318.

11. Xu, C., Lopez, R., Mehlman, E., Regier, J., Jordan, M. I., and Yosef, N., Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.*, 2021, **17**, e9620.
12. Steinert, E. M., Schenkel, J. M., Fraser, K. A., et al., Quantifying Memory CD8 T Cells Reveals Regionalization of Immunosurveillance. *Cell*, 2015, **161**, 737–749.
13. Farooqui, M., Bohrer, L. R., Brady, N. J., et al., Epiregulin contributes to breast tumorigenesis through regulating matrix metalloproteinase 1 and promoting cell survival. *Mol. Cancer 2015 141*, 2015, **14**, 1–13.
14. Riese, D. J., II, and Cullum, R. L., Epiregulin: Roles in Normal Physiology and Cancer. *Semin. Cell Dev. Biol.*, 2014, **0**, 49.
15. Godfrey, D. I., Koay, H.-F., McCluskey, J., and Gherardin, N. A., The biology and functional importance of MAIT cells. *Nat. Immunol. 2019 209*, 2019, **20**, 1110–1128.
16. Gillich, A., Zhang, F., Farmer, C. G., et al., Capillary cell-type specialization in the alveolus. *Nat. 2020 5867831*, 2020, **586**, 785–789.
17. O'Leary, N. A., Wright, M. W., Brister, J. R., et al., Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, 2016, **44**, D733–D745.
18. Dehghannasiri, R., Olivieri, J. E., and Salzman, J., Specific splice junction detection in single cells with SICILIAN. *bioRxiv*, 2020, 2020.04.14.041905.
19. Pertea, M., Shumate, A., Pertea, G., et al., CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

extensive transcriptional noise. *Genome Biol.*, 2018, **19**.

20. Wan, Y., Anastasakis, D. G., Rodriguez, J., et al., Dynamic imaging of nascent RNA reveals general principles of transcription dynamics and stochastic splice site selection. *Cell*, 2021, **184**, 2878-2895.e20.
21. Olivieri, J. E., Dehghannasiri, R., and Salzman, J., The SpliZ generalizes “Percent Spliced In” to reveal regulated splicing at single-cell resolution. *bioRxiv*, 2021, 2020.11.10.377572.
22. Sugi, H., *Muscle Contraction and Cell Motility Muscle Contract. Cell Motil.*, Jenny Stanford Publishing, 2016.
23. Fisher, S. A., Vascular smooth muscle phenotypic diversity and function. *https://doi.org/10.1152/physiolgenomics.00111.2010*, 2010, **42 A**, 169–187.
24. Majeti, R., Chao, M. P., Alizadeh, A. A., et al., CD47 Is an Adverse Prognostic Factor and Therapeutic Antibody Target on Human Acute Myeloid Leukemia Stem Cells. *Cell*, 2009, **138**, 286–299.
25. Hayat, S. M. G., Bianconi, V., Pirro, M., Jaafari, M. R., Hatamipour, M., and Sahebkar, A., CD47: role in the immune system and application to cancer therapy. *Cell. Oncol. 2019 431*, 2019, **43**, 19–30.
26. Gehart, H. and Clevers, H., Tales from the crypt: new insights into intestinal stem cells. *Nat. Rev. Gastroenterol. Hepatol.*, Nat Rev Gastroenterol Hepatol, 2019, January 1, pp. 19–34.
27. Manno, G. La, Soldatov, R., Zeisel, A., et al., RNA velocity of single cells. *Nature*, 2018, **560**, 494.
28. Bergen, V., Lange, M., Peidli, S., Wolf, F. A., and Theis, F. J., Generalizing RNA

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.*, 2020, **38**, 1408–1414.

29. Murphy, A. C. H. and Young, P. W., The actinin family of actin cross-linking proteins - A genetic perspective. *Cell Biosci.*, Cell Biosci, 2015, August 25.

30. Li, R., Li, X., Hagood, J., Zhu, M.-S., and Sun, X., Myofibroblast contraction is essential for generating and regenerating the gas-exchange surface. *J. Clin. Invest.*, 2020, **130**, 2859–2871.

31. Mowat, A. M. and Agace, W. W., Regional specialization within the intestinal immune system. *Nat. Rev. Immunol. 2014 1410*, 2014, **14**, 667–685.

32. Halfvarson, J., Brislawn, C. J., Lamendella, R., et al., Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat. Microbiol.*, 2017, **2**.

33. Donaldson, G. P., Lee, S. M., and Mazmanian, S. K., Gut biogeography of the bacterial microbiota. *Nat. Rev. Microbiol.*, Nat Rev Microbiol, 2015, December 16, pp. 20–32.

34. James, K. R., Gomes, T., Elmentaite, R., et al., Distinct microbial and immune niches of the human colon. *Nat. Immunol.*, 2020, **21**, 343–353.

35. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, nmeth.2639 (2013).

36. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

37. Griffiths, J. A., Richard, A. C., Bach, K., Lun, A. T. L. & Marioni, J. C. Detection and removal of barcode swapping in single-cell RNA-seq data. *Nat. Commun.* **9**, 2667 (2018).

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

38. Yang, S. *et al.* Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biol.* **21**, 57 (2020).

39. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1 11 (2018).

40. Gayoso, A. *et al. scvi-tools: a library for deep probabilistic analysis of single-cell omics data*. http://biorxiv.org/lookup/doi/10.1101/2021.04.28.441833 (2021) doi:10.1101/2021.04.28.441833.

41. Diehl, A. D. *et al.* The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. *J. Biomed. Semant.* **7**, 44 (2016).

42. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

43. Wang, S. *et al.* Unifying single-cell annotations based on the Cell Ontology. *bioRxiv* 810234 (2020) doi:10.1101/810234.

44. Polański, K. *et al.* BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics* (2019) doi:10.1093/bioinformatics/btz625.

45. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).

46. Sofroniew, N. *et al. napari/napari: 0.4.10rc0*. (Zenodo, 2021). doi:10.5281/ZENODO.4968798.

47. Graham, S. *et al.* Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med. Image Anal.* **58**, 101563 (2019).

48. Core Team, R. R: A Language and Environment for Statistical Computing. (2017).

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

49. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* **13**, 581–583 (2016).

50. Yilmaz, P. *et al.* The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res.* **42**, D643–D648 (2014).

51. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).

52. McMurdie, P. J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE* **8**, e61217 (2013).

53. Whitfield, M. L. *et al.* Identification of Genes Periodically Expressed in the Human Cell Cycle and Their Expression in Tumors. *Mol. Biol. Cell* **13**, 1977–2000 (2002).

54. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).

55. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).

56. Santos, A., Wernersson, R. & Jensen, L. J. Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. *Nucleic Acids Res.* **43**, D1140–D1144 (2015).

57. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).

58. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).

59. Butler, A et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotech* 36, pages 411–420 (2019)

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

60. van den Brink, S. C. *et al.* Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nat. Methods* **14**, 935–936 (2017).

61. Ferreira, P. G. *et al.* The effects of death and post-mortem cold ischemia on human tissue transcriptomes. *Nat. Commun.* **9**, 490 (2018).

62. Massoni-Badosa, R. *et al.* Sampling time-dependent artifacts in single-cell genomics studies. *Genome Biol.* **21**, 112 (2020).

63. van Velthoven, C. T. J., de Morree, A., Egner, I. M., Brett, J. O. & Rando, T. A. Transcriptional Profiling of Quiescent Muscle Stem Cells In Vivo. *Cell Rep.* **21**, 1994–2004 (2017).

64. Savage, A. K. *et al. Multimodal Analysis for Human ex vivo Studies Shows Extensive Molecular Changes from Delays in Blood Processing*. http://biorxiv.org/lookup/doi/10.1101/2020.10.18.344663 (2020) doi:10.1101/2020.10.18.344663.

65. Stubbington, M. J. T. *et al.* T cell fate and clonality inference from single-cell transcriptomes. *Nat. Methods* **13**, nmeth.3800 (2016).

66. Lindeman, I. *et al.* BraCeR: B-cell-receptor reconstruction and clonality inference from single-cell RNA-seq. *Nat. Methods* **15**, 563–565 (2018).

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

68. Ye, J., Ma, N., Madden, T. L. & Ostell, J. M. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* **41**, W34–W40 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

69. Sturm, G. *et al.* Scirpy: a Scanpy extension for analyzing single-cell T-cell receptor-sequencing data. *Bioinformatics* **36**, 4817–4818 (2020).

70. Peixoto, T. P. The graph-tool python library. 136152697 Bytes (2017) doi:10.6084/M9.FIGSHARE.1164194.

71. Waskom, M. seaborn: statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).

Acknowledgements: We express our gratitude and thanks to donor WEM and his family, as well as all of the anonymous organ and tissue donors and their families for giving both the gift of life and the gift of knowledge by their generous donations. We also thank Donor Network West for their cooperation in this project, Sandy Schmid for a close reading of the manuscript, and Bruno Tojo for the original artwork in Figure 1. This project has been made possible in part by grant number 2019-203354 from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation, and by support from the Chan Zuckerberg Biohub.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 Morri[2], 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],

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 Qi[18], Poorvi Rao[20], Hayley Raquer-McKay[16], Nicholas Schaum[18,21], Bronwyn Scott[31], Bobak Seddighzadeh[38], Joe Segal[20], Sushmita Sen[29], 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]

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 Raquer-McKay[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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure Legends:

Figure 1. Overview of Tabula Sapiens. The Tabula Sapiens was constructed with data from 15 human donors. Donors 1, 2, 7 and 14 contributed the largest number of tissues each, and the number of cells from each tissue is indicated by the size of each circle. Tissue contributions from additional donors are shown as well, and the total number of cells for each organ are shown in the final column.

Figure 2. Comparison of single cell transcriptomics with conventional histology. A. H&E stained image used for histology of the colon, with compartments (solid, colored lines) and individual cell types (dashed black ellipses) identified by the pathologists. **B**. Coarse cell type representation of donor 2 as morphologically estimated by pathologists across several tissues, ordered by increasing heterogeneity of the tissue. **C**. Comparison of coarse cell type abundance between histology-based estimates and sequencing for donors 1, 2 and 14. Data is across all tissues analyzed and is binned into four levels of abundance as estimated by pathologists.

Figure 3. Analysis of immune and endothelial cell types shared across tissues. A. Illustration of clonal distribution of T cells across multiple tissues. The majority of T cell clones are found in multiple tissues and represent a variety of T cell subtypes. **B**. Prevalence of B cell isotypes across tissues, ordered by decreasing abundance of IgA. **C**. Expression level of tissue specific endothelial markers, shown as violin plots. Many of the markers are highly tissue specific.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 4. Alternative splicing analysis. A,B. The sixth exon in MYL6 is skipped at different proportions in different compartments. Cells in the immune and epithelial compartments tend to skip the exon, whereas cells in the endothelial and stromal compartments tend to include the exon. Boxes are grouped by compartment and colored by tissue. The fraction of junctional reads that include exon 6 was calculated for each cell with more than 10 reads mapping to the exon skipping event. Only shared cell types with more than 10 cells with spliced reads mapping to MYL6 are shown. **C,D**. For CD47, epithelial cells tend to use closer exons to the 108047292 5' splice site compared to immune and stromal cells. Boxes are grouped by compartment and colored by tissue. The four splice sites corresponding to this 5' splice site were ranked (as in the legend), and the average splice site rank was calculated for each cell with more than one read corresponding to the splice site. Only shared cell types with more than 5 cells with spliced reads mapping to CD47 are shown.

Figure 5. Dynamic changes in cell state. A. Cell types ordered by magnitude of cell cycling index, with the most highly proliferative at the top and quiescent cells at the bottom of the list. **B**. RNA velocity analysis demonstrating mesenchymal to myofibroblast transition in the bladder. **C**. Latent time analysis of the mesenchymal to myofibroblast transition in the bladder demonstrating highly stereotyped changes in gene expression trajectory.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 6. High resolution view highlights patchiness of the gut microbiome. A,B. Schematic (A) and photo of the colon from donor TSP 2 (B) with numbers 1-5 representing microbiota sampling locations. **C**. Relative abundances (top) and richness (number of observed species, middle) at the family level in each sampling location, as determined by 16S rRNA sequencing. The Shannon diversity, a metric of evenness, mimics richness. Variability was higher in the duodenum, jejunum, and ileum as compared with the ascending and sigmoid colon. **D**. A Sankey diagram showing the inflow and outflow of microbial species from each section of the gastrointestinal tract. The stacked bar for each gastrointestinal section represents the number of observed species in each family as the union of all sampling locations for that section. The stacked bar flowing out represents gastrointestinal species not found in the subsequent section and the stacked bar flowing into each gastrointestinal section represents the species not found in the previous section.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Figure Legends:

Supplementary Figure 1. Data processing workflow to build the reference dataset. Schematic representation of the preprocessing and annotation steps taken to build Tabula Sapiens.

Supplementary Figure 2. UMAP overview of key metadata variables in Tabula Sapiens, colored by organ/tissue (**A**). donor (**B**), compartment (**C**), method (**D**) and sex (**E**).

Supplementary Figure 3. Droplet (10x) sequencing statistics for genes detected. Box plot of the number of genes detected per cell for each organ and donor

Supplementary Figure 4. Droplet (10x) sequencing statistics for number of UMIs. Box plot of the number of UMIs per cell (log-scale) for each organ and donor.

Supplementary Figure 5. Smartseq2 sequencing statistics for genes detected. Box plot of the number of genes detected per cell for each organ and donor

Supplementary Figure 6. Smartseq2 sequencing statistics for number of counts. Box plot of the number of reads per cell (log-scale) for each organ and donor.

Supplementary Figure 7. Tabula Sapiens metadata summary. Summary of cell compartment distribution across tissues and donors as measured by 10X and Smartseq2 sequencing.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Figure 8. Comparison of cell-type frequency with literature. Prior to single-cell analysis the literature was searched to establish the expected cell types and relative abundance in each tissue. **A**. The frequency of lung cell-types identified are compared to the in-depth curated literature citations of a recent lung-specific cell atlas. The plot shows the effects of dissociation artefacts and compartment enrichment. **B**. Bladder cell-type frequency vs estimates from literature and gene expression data. Myocytes have been removed from the plot as these were trimmed from samples. **C**. Muscle cell types are linear with estimates from the literature.

Supplementary Figure 9. Tissue compartment compositions for TSP14. Spatial heterogeneity was used to sort the tissue bar plot showing the relative abundances of cell types per functional compartment in each tissue for the droplet dataset of TSP14.

Supplementary Figure 10. **Immune cell lineage and gene expression across tissues A**. Tissue-level correlation of gene expression within macrophage for tissue-specific genes. **B**. Empirical cumulative distribution plots of gene expression for selected macrophage genes. Some genes are highly specific to one tissue, others are specific to multiple tissues. **C**. Graph-based illustration of clonal distribution of T cells across multiple tissues. The majority of T cell clones are found in multiple tissues **E**. Somatic hypermutation levels in the V-genes of B cells sampled from different tissues in Tabula Sapiens.

Supplementary Figure 11. Tissue specific gene expression patterns in endothelial cells. **A,B**. UMAP depicting endothelial cells across tissues (**A**) and donors (**B**). **C,D**. UMAP depicting expression distribution (scale is lnCPM) of PLVAP (**C**) and EDNRB (**D**) across gCap and aCap populations, respectively. **E,F**. UMAP depicting expression distribution (scale is lnCPM) of MSX1 (**E**) and CYP1B1 (**F**) in subpopulations of muscle ECs.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Figure 12. Alternative splicing analysis A. Splice junctions are counted in categories based on annotation status. **B,C**. Pie charts show the breakdown of the number of junctions by annotation category (**B**) and the number of reads (**C**) per annotation category. Although unannotated junctions make up the majority of unique junctions, annotated junctions make up the majority of total reads. **D**. Transcript annotation of CD47. The analyzed 5' splice site is marked in blue and the four alternative 3' splice sites are marked in red.

Supplementary Figure 13. Cell state dynamics. A. Cell cycle indices for various cell types in the small intestine. **B**. Immunohistochemical stain for Ki67 of the jejunum demonstrates high proliferation amongst crypt stem cells and transient amplifying cells, situated just above the base of the crypts. Schematic shown is modified from (Gehart & Clevers, 2019). **C**. Leiden clustering of transient amplifying cells with top differentially expressed cell cycle marker genes. **D**. UMAP clustering of macrophages in donor TSP2, from different tissues, **E,F**. Same UMAP color-coded based on the binary assignment of cycling and non-cycling (E) and CDKN1A expression (F). **G**. cycling indices of TSP2 macrophages across different tissues. **H**. (left) MYLK and ACTN1 expression as a function of latent time across TSP1 and TSP2; (right) **I.** A random sampling of dynamical genes shared between TSP1 and TSP2, showing the expression of dynamical genes as a function of latent time during the transitioning of mesenchymal cells to myofibroblasts.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Tables:

Supplementary Table 1. Donor summaries.

Supplementary Table 2. Dataset summary statistics.

Supplementary Table 3. Literature estimates for cell types.

Supplementary Table 4. Tabula Sapiens provisional cell ontology. Table of cell type label to its parent cell type label(s) in the reference ontology. Cell types with asterisk denotes missing cell types in the public cell ontology that were added.

Supplementary Table 5. Genes affected by dissociation.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

Organ and tissue procurement

Donated organs and tissues were procured at various hospital locations in the Northern California region through collaboration with a not-for-profit organization, Donor Network West (DNW, San Ramon, CA, USA). DNW is a federally mandated organ procurement organization for Northern California. Recovery of non-transplantable organ and tissue was considered for research studies only after obtaining records of first-person authorization (i.e., donor's consent during his/her DMV registrations) and/or consent from the family members of the donor. Each tissue was collected, and transported on ice, as quickly as possible to preserve cell viability. A private courier service was used to keep the time between organ procurement and initial tissue preparation to less than one hour. Single cell suspensions from each organ were prepared in tissue expert laboratories at Stanford and UCSF. The research protocol was approved by the DNW's internal ethics committee (Research project STAN-19-104) and the medical advisory board, as well as by the Institutional Review Board at Stanford University which determined that this project does not meet the definition of human subject research as defined in federal regulations 45 CFR 46.102 or 21 CFR 50.3

Bladder - Initial tissue preparation protocol

Tissue was collected around the area of the mucosa and submucosa in the bladder dome, trying to avoid the muscular tissue surrounding the region of interest. The collected tissue was prepared into a single cell solution via enzymatic digestion. The initial tissue volume after removal of the thick muscle layers surrounding the mucosa was approximately 1 cm^3. Tissue was minced and digested first using the Digest A buffer (Collagenase type IV1875 U (~75 mg, depending on batch), 100 µL of P 188 (10%, 100x), 10 µL of 1M CaCl2, 500 µL of 20% PVP, 25 µL of DNase (4U/µL), 9.5 mL of M199-HEPES) in a 50-mL Falcon tube. The first digestion in Digestion Buffer A was performed at 37°C for 1 h. The second digestion step was performed using Digestion Buffer B (1 mL of 10x TrypLE, 100 µL of P 188 (10%, 100x), 100 µL of 0.5 M EGTA, 500 µL of 20% PVP, 25 µL of DNase (4U/µL), 8.7 mL of M199-HEPES). Before digestion, Digestion Buffer B was prewarmed for 5 minutes at 37°C. Cells in Digestion Buffer A were pelleted at 500g for 5 minutes using a swinging bucket rotor. Supernatant was removed, leaving only 100 µL in the Falcon tube. 500 µL of pre-warmed Digestion Buffer B was added to the tissue, and the samples were titrated 15 times using a wide-bore pipet tip. A final volume of 9.5 mL of Digestion Buffer B was added and tissue was incubated at 37°C for 30 minutes (200 rpm). Samples were kept on ice from this point on, using chilled buffers and a chilled centrifuge (set at 4°C). Cell suspensions were centrifuged at 500g for 5 minutes at 4°C using a swing bucket rotor to obtain a visible pellet. Supernatant was removed, leaving a total volume of 200 µL. 500 mL of M199 were added to the 200 µL of cells previously obtained and titrated 15 times using a wide bore pipet tip. In the meantime, a 50-mL Falcon tube topped with a 40-µm cell strainer was put on ice. 5 mL of M199 were added to the cells and passed through the cell strainer into the Falcon tube. An additional 10 mL of M199 were added to the digestion tube; the tube was inverted to wash away any remaining cells, and the suspension was passed again through the strainer. The strainer was washed with 10 mL of FACS buffer. The cell suspension was again centrifuged at 500g for 5 minutes at

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4°C. Supernatant was removed, leaving a total volume of 100 µL. Cells contained in the 100 µL were resuspended using 1 mL of RBC lysis buffer and incubated at room temperature for 5 minutes. 10 mL of FACS buffer were added to the cells plus RBS lysis buffer mix, and centrifuged at 500g for 5 minutes at 4°C. Supernatant was removed, leaving behind a total volume of 100 µL. 900 µL of FACS buffer were added to the mix, which was titrated 15 times using a wide bore pipet. Cell concentration was then measured using a hemocytometer. The final cell suspension was divided into two 500 µL samples to be further prepared for 10x Genomics and FACS sorting/SS2.

Bladder – 10x Genomics sample preparation

The 500 µL sample of cell suspension designated for 10x Genomics was centrifuged at 500g for 5 minutes at 4°C. In the meantime, a mix of CD31/CD45 antibodies was prepared for subsequent MACS selection. The antibody mix was prepared using 100 µL of MACS buffer, 2 µL of CD45 beads (Miltenyi 130-045-801), 1 µL of CD31 beads (Miltenyi 130-091-935), and 1 µL of FcBlock beads (Miltenyi 130-059-901). After centrifugation, the supernatant was removed and cells were resuspended in the just-prepared antibody mix for 20 minutes on ice. After incubation, cells were centrifuged at 500g for 5 minutes at 4°C. After centrifugation, cells were resuspended in 500 µL of MACS buffer using a wide bore pipet. Cell suspension was again centrifuged at 500g for 5 minutes at 4°C. These two steps were repeated for a total of two cell washes. The MACS column was equilibrated by adding 1 mL of MACS buffer with the column on the magnet; a 15-mL Falcon collection tube labelled CD31-/CD45- was placed on ice below the columns. The cell suspension was applied to the column immediately after 1 mL of MACS buffer was added, being careful to allow the cell suspension and the buffer to enter the column and collect the flow through. An additional 5 mL of MACS buffer was added and allowed to pass through the columns, collecting the flow-through. This process was repeated a second time. The tube was set aside, now containing only CD31-/CD45- negative cells. The column was removed from the magnet and a new 15-mL Falcon tube labeled as CD31+/CD45+ was placed. 5 mL of MACS buffer were added to the column and allowed to run through the column. The sample was plunged to recover any remaining material. Both Falcon tubes were centrifuged at 500g for 5 minutes at 4°C. During this centrifugation, the EPCAM MACS beads were prepared for the second MACS selection step. EPCAM mix was prepared using 100 µL of MACS buffer, 2 µL of EPCAM magnetic beads, and 1 µL of FcBlock magnetic beads. Upon completion of centrifugation, the supernatant was removed. For the Falcon tube with CD31-/CD45- cells, the cell pellet was resuspended in the MACS/EPCAM mix previously prepared and incubated on ice for 20 minutes. After incubation, cells were washed twice using the same protocol as for the first MACS step. A new MACS column was placed on the magnetic holder and a Falcon tube was labeled with CD31-/CD45-/EPCAM-. The column was equilibrated using 1 mL of MACS buffer, and cell suspension was added to the column followed immediately by addition of 1 mL of MACS buffer. The cell suspension and the buffer were allowed to run through the column. 5 mL of MACS buffer were added to the column, the flow was collected, and another 5 mL was added. Once column flow completed, the CD31-/CD45-/EPCAM- Falcon was set aside. The column was removed from the magnet and a new 15-mL Falcon tube labelled as CD31-/CD45-/EPCAM+. 5 mL of MACS buffer was added to the column and the flow was collected. The sample was plunged to collect any

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

remaining material stuck in the column. For all populations (CD31+/CD45+, CD31-/CD45-/EPCAM-, CD31-/CD45-/EPCAM+), the collections were centrifuged at 500g for 5 minutes at 4°C and then the supernatants were aspirated and resuspended in 200µL FACS buffer. Cell concentrations and viability were measured using a hemocytometer. Cell concentrations were adjusted to 10^6 cells/mL. Different cell populations were combined 1:1:1 in the same tube to balance the different compartments and were ready for 10x downstream pipeline.

Bladder – FACS/SS2 sample preparation

The second 500 µL aliquot of cell suspension was used for the SS2 protocol. Cell concentration was adjusted to 2-3×10^6 cells per mL. The following antibodies were used at a 1:20 dilution in FACS buffer: CD45-FITC, CD31-PE, and EPCAM-Cy7. Cells were stained on ice for 20 minutes in 70 µL (single antibodies) or 200 µL (all antibodies). After incubation, 1 mL FACS buffer was added to the cells, and cells were pelleted at 500g for 5 minutes at 4°C in a swinging bucket centrifuge. Finally, supernatants were aspirated and cells were resuspended in 1 mL of FACS buffer using a wide bore pipet. Final single cell solutions were passed through a 40-µm cell strainer into a FACS tube to generate samples ready for FACS sorting.

Blood – Initial tissue preparation protocol

4.5 mL of blood was mixed with 45 mL of FACS buffer (PBS plus 2% FBS) and 0.5 mL of DNase I at room temperature (initial DNase I concentration was 5.5 mg/mL, final concentration was 0.05 mg/mL). 15 mL of Ficoll Histopaque-1119 were added to two empty 50-mL Falcon tubes. 25 mL of blood/FACS buffer mixture were added to each Ficoll-filled Falcon tube, tilting the tube and pipetting on its side to prevent the blood from mixing with the Ficoll. Both tubes were centrifuged at 400g for 30 minutes at room temperature, with the centrifuge's brakes off. After centrifugation, the tubes were inspected to check that all blood cell layers were well separated. Starting from the bottom, the following layers were identified: erythrocytes, Ficoll solution, buffy coat with cells (white color), and plasma. The buffy coats were gently removed from each tube and transferred into a new 50-mL Falcon tube. 30 mL of cold (4°C) FACS buffer were added to each buffy coat. The Falcon tubes were centrifuged at 300g for 5 minutes at 4°C. After centrifugation, an additional 5 mL of cold FACS buffer were added. Cells were counted using a hemocytometer. For whole blood, 1 mL of blood was added to 10 mL of ACK lysis buffer for 10 minutes at room temperature. The cells were then spun down at 4ºC, 400g for 5 minutes. After the ACK lysis step, cells were washed with 10 mL of ice-cold PBS and spun down at 4°C, 450 g for 10 minutes. Cells were then resuspended in 5 mL of FACS buffer and counted.

Blood - FACS/SS2 sample preparation

5 million cells were stained for Lineage-PE (CD3, CD19, CD20, CD335 and CD66b) and monocytes (CD14- and CD16-APC). Sytox blue was added and live, single cells were sorted based on Lineage+Monocyte-, Lineage-Monocyte+ and Lineage-Monocyte-.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Blood – 10x Genomics sample preparation

10 million cells were stained with purified mouse anti-human CD3, CD19, CD20, CD335, and CD14 for 30 minutes on ice and washed twice with media. Cells were then negatively selected by adding pan anti-mouse Dynabeads (ThermoFisher 11041) using 4 Dynabeads per target cell following manufacturer's instructions. Cell suspensions were adjusted to 10^6 cells/mL prior to running the sample in the 10x Genomics controller. For each donor, one sample consisted of negatively selected cells and one sample consisted of 25% whole blood after red blood cell removal and 75% PBMCs after density gradient centrifugation with Ficoll-Paque PLUS.

Bone Marrow

Human Vertebral Bodies - Initial Tissue preparation protocol

The vertebral bodies (VB) were wrapped in a cloth and shipped to Stanford University on ice. Upon arrival, the VBs were first cleaned using chisels, to remove any attached connective tissue and fat. To rinse off any remaining cells attached to the exterior of the VB, the VBs were then transferred to a plastic nalgene containing 50 mL RPMI + 10% FBS and tumbled in a bone marrow tumbler for 20 minutes at room temperature. The rinsing medium was discarded and the VBs were removed from the tumbler, transferred to a large sterile plastic Petri dish, and cut into ½ inch by ½ inch pieces using bone cutting forceps. The bone marrow pieces were transferred into a plastic nalgene, to which 100mL of RPMI + 10% FBS was added. The nalgene was returned to the bone marrow tumbler and tumbled for 30 minutes at room temperature. The solution containing the VB was then passed through a 100 µm strainer into 50 mL falcon tubes. Multiple strainers were used in case of clogs. After straining, the cells were centrifuged and pelleted at 330g 4°C for 5 minutes. Cells were resuspended in 1x Erythrocyte Lysis buffer, kept 5 minutes on ice and mixed. Cells were then re-centrifuged for 5 minutes at 330g at 4°C without brakes, in order to remove plasma and lysed red blood cells. Cells were then ready to count. 10 million bone marrow cells were stained using an immune lineagePE cocktail containing CD3-PE, CD4-PE, CD56-PE, CD11b-PE, and CD14-PE, and subsequent MACS with anti-PE microbeads was performed. One tube of immune lineage positive and one tube of lineage negative cell were prepared for the 10x. For smartseq2, cells were stained with anti-CD38-APC, anti-CD34-FITC, and sytox blue for 30 minutes at 4°C and washed twice.

Eye - Initial Tissue preparation protocol

The eye and extraocular components were enucleated from the donor and transported to Stanford University on ice. The eye was carefully dissected using forceps and scissors and the different components rinsed with PBS. The central cornea, limbus and conjunctiva were dissected out first, lightly minced using blades and incubated in 1U/mL Dispase in DMEM/F12 (Fisher Scientific, NC9995391) for 2 hours at 37°C with shaking. The sclera, iris, lens and optic nerve were dissected out next, minced using blades and also digested in 1U/mL Dispase in DMEM/F12 (Fisher Scientific, NC9995391) for 2 hours at 37°C with shaking. For retinal dissociation, papain solution was prepared using 0.8µl Papain (100mg/mL) (Worthington Biochemicals, LS003119), 4µl L-cysteine (250mM), 0.2µl EDTA (0.5M, pH 8) and 395µl HBSS(Thermo Fisher Scientific, 14025076). Papain solution was incubated at 37°C for 10 minutes to ensure activation. The retina and RPE were peeled off from the optic cup, lightly minced and incubated in the papain solution at

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

37°C for 15 minutes. Eyelids were minced in 0.25% collagenase (Stemcell Technologies, NC9952277) and lacrimal glands minced in 0.125% collagenase (Stemcell Technologies, NC9952277) and incubated at 37°C for 1 h and 0.5 h, respectively. Orbital fat was incubated in 0.075% collagenase (Stemcell Technologies, NC9952277) for 90 minutes and extraocular muscle was incubated in 0.2% collagenase + 8mM $CaCl_2$ at 37°C for 1 h. After digestion, the tissues were further minced in 0.05% Trypsin EDTA (Thermo Fisher Scientific, 25300054) for 5 minutes. The enzymatic reaction was stopped with equal volumes of DMEM + 10% FBS. The cells were filtered through a Falcon 40 µm cell strainer (Corning, 352340) and centrifuged at 300g for 5 minutes. Cells were washed 3 times with DMEM + 10% FBS and centrifuged. Cells were counted using a hemocytometer and resuspended at a concentration of 1000 cells/µL in DMEM + 10% FBS for 10x Genomics. For smartseq-2 the cells were labeled with EPCAM-PE, CD45-FITC, CD31-APC and all three antibodies for 30 minutes. Cells were sorted according to the gating strategy live/CD45-/EPCAM+ for epithelial cells, live/CD45+/EPCAM- for immune cells, live/CD45-/EPCAM-/CD31+ for endothelial cells and live/CD45-/EPCAM-/CD31- for stromal or neuronal cells. The cells from different eye tissues were combined at 1:1 ratio and pooled for 10x.

Fat - Initial Tissue preparation protocol

Fat tissues were dissected out and minced before digestion in 760 U/mL Collagenase II (Worthington LS004177) and Dispase II (Gibco 17105-041, 1U/mL), shaking for 30-60 minutes at 37°C. Cells were then filtered consecutively through 100 µm (Falcon 352360) and 40 µm (Falcon 21 352340) strainers on ice and washed with cold F-10/Ham's medium containing 10% horse serum (Invitrogen 16050114). One aliquot of cells was resuspended to 1000 cells/µL for cell capture in microfluidic droplets with the 10x Genomics platform. The remaining cells were stained with 1:100 CD31-APC (Biolegend 303116), 1:500 CD45-FITC (Biolegend 304038), and 1:50 CD200-PE (BD Biosciences 561762) for MAT, or 1:50 CD10-PE (BD Biosciences 561002) for SCAT for 30 minutes at 4°C in F-10/Ham's with 10% horse serum and washed. 1:1000 SytoxBlue (Invitrogen S34857) was added immediately prior to sorting. Cells were sorted into 3 bins: mesenchymal progenitors (CD200+ or CD10+, CD31- , CD45-), immune cells (CD45+, CD31-), and endothelial cells (CD31+, CD45-).

Heart – Initial Tissue Preparation Protocol

Small 1cm x 1cm tissue chunks were dissected from the left and right atrial (appendage) and left and right ventricular (free wall) myocardium. Tissues were then dissociated into single cells in a microcentrifuge tube with 400 µL of 0.25% trypsin and incubated at 37°C for 10 minutes. Subsequently, 1.6 mL of collagenase A/B (10 mg/mL, Roche) and 20% FBS serum in HBSS was added to the microcentrifuge tubes and tissue samples were returned to 37°C water bath for an additional 20 minutes with intermittent tissue disruption by manually pipetting using a 1000 µL pipette tip. Cells were then filtered through a 40 µM filter (Falcon) prior to spinning down at 1000 RPM for 5 minutes, supernatant was removed, and cells were washed in 20% FBS in HBSS three times. Red blood cells were lysed with ACK lysis buffer for 5 minutes prior to final wash step. Atrial cells were pooled into one sample and ventricular cells into a separate sample. After the final wash and

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

supernatant removal, cells were resuspended to a concentration of around 600 cell/µL with 0.04% FBS/HBSS solution for processing on the 10x platform.

Intestine – Initial tissue preparation protocol

The following parts of the small and large intestines were collected from the donor: duodenum, jejunum, ileum, ascending colon and sigmoid colon. The intestine tissues were rinsed with 30 mL of ice-cold PBS. Intestines were then transferred to a new petri dish, opened longitudinally and all intestinal contents were flushed out using ice-cold PBS. The intestine tissues were minced into small pieces (2-3 mm in size) with a clean blade. After this, collagenase III (200 unit/mL) and DNAse I (100 unit/mL) in 10 mL of digestion medium (DMEM/F12, 1x PSA, 10mM HEPES, 1mM sodium pyruvate) was added. Tissues were digested at 37°C for 90-120 minutes with pipetting every 15 minutes using a 10 mL pipet. Digested tissue was transferred to a 50 mL Falcon tube. FACS buffer (HBSS, 2% FBS, HEPES, 1% PSA) was added to get to 30 mL of total volume. The cells were spun down at 500g, for 5 minutes at 4°C and the supernatant was discarded. Red blood cells were lysed with ACK lysis buffer for 5 minutes. FACS buffer was added to stop the reaction. The reaction was spun down in the centrifuge at 500 g, for 5 minutes at 4°C and the supernatant was discarded. The cells were resuspended with 10 mL of FACS buffer with 100unit/mL DNaseI and passed through a 40 µm cell strainer.

Kidney - Initial Tissue preparation protocol

For the kidney, a wedge biopsy was performed obtaining a tissue sample of the size of 1cm x 1cm x 2mm). Tissue was placed in a petri dish on ice and minced finely using a razor blade. The minced tissue was placed into a Miltenyi C tube with 10mL of cold liberase TL buffer. The tissue was then placed into the GentleMacs; after that the tissue was put into a shaker at 37°C for 20 minutes. Tissue was placed for a second time into the GentleMacs. After that, 2 mL of FBS were added to stop the digestion and tissue was placed on ice. Tissue solution was passed through a 70 µm SmartStrainer and the strainer was washed with 15 mL cold RPMI to collect any remaining cells. Cells were centrifuged at 300g for 7 minutes at 4°C. Supernatant was removed and the cells were first resuspended in a minimal residual volume and then 10 mL of RPMI were added. Cells were counted and ready for smartseq2 and 10x protocols.

Liver – Initial Tissue preparation protocol

The freshly explanted human livers were kept on ice in Belzer UW Cold Storage Solution (BTL company, Northbrook, IL). Total cell suspensions were obtained via mechanical enzymatic digestion, between 3-5 hours after surgery. The liver was dissociated by mincing the tissue into 1-2mm squares, and incubating in Liver Perfusion Medium (ThermoFisher Scientific, Waltham, MA) for 15 minutes at 37°C with rotation. After washing in Dulbecco's PBS (ThermoFisher Scientific), tissues were incubated with HBSS (ThermoFisher Scientific) supplemented with HEPES (ThermoFisher Scientific) and Collagenase type IV (Worthington Biochemical, Lakewood, NJ; 600 U/mL) for 30 minutes at 37°C with rotation. Remaining pieces of tissue were further dissociated by pipetting 10 times through a 25 mL serological pipette, and single cells were separated from clumps using a 70 µm strainer (Fisher Scientific, Hampton, NH). After lysis of red blood cells with ACK RBC Lysing Buffer (Fisher Scientific), Hepatocytes were separated from the non-

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

parenchymal cells with slow speed centrifugation at 30g. Each cell fraction was further washed using Williams E medium supplemented with Glutamax, Non-Essential Amino Acid, HEPES, and Pen-Strep (ThermoFisher). Cells were counted using a LUNA Automated Cell Counter (Logos Biosystems, South Korea) and processed for flow-cytometry sorting and scRNAseq using SmartSeq2. For 10x single cell capture, hepatocyte and non-parenchymal cell fractions were mixed 1:1.

Lung - Initial Tissue Preparation Protocol

Upon arrival on ice, lung tissues were separated from the bronchus(proximal), mid-bronchial region (medial), and periphery (distal), which were processed identically and separately. The tissue was then chopped roughly with scissors on the side of a gentleMACS type C tube and 12 mL of Liberase medium (400 µg/mL Liberase DL (Sigma 5466202001) and 100 µg/mL elastase (Worthington LS006365) in RPMI (Gibco 72400120)) were added to the tube. Tubes with tissue were inserted into the gentleMACS Dissociator (Miltenyi 130-093-235) and the program “m_lung_01” was run. Tubes were then placed upside down in a styrofoam box and incubated at 37°C for 30 minutes. After this first incubation, tubes were inserted again into the gentleMACS dissociator and the program “m_lung_02” was run. The Liberase medium was then neutralized using 12 mL of 5% FBS. Cells were spun down at 500g for 10 minutes at 4°C, and later resuspended in 10 mL of ice cold 5% FBS. Samples were then placed at 4 °C for the remainder of the protocol. A 100-µm filter was placed on top of a clean 50-mL Falcon tube and cell suspension was passed through the filter. Tissue chunks on top of the filter were smashed using the plunger of a 3-mL syringe. The gentleMACS tube was then washed using 5 mL of 5% FBS and the solution was pipetted onto the filter. Cells were then spun down at 500g for 10 minutes and the cell pellet was resuspended in 1 mL of ACK RBC lysis buffer for 2 minutes. The ACK was then neutralized by adding 5 mL of 5% FBS. Cells were pelleted (500g, 10 minutes), resuspended in 6 mL 5% FBS in PBS, filtered through a 70-µm strainer (Fisherbrand 22363548), pelleted again, and resuspended using 200 µL of magnetic activated cell sorting (MACS) buffer (0.5% BSA, 2 mM EDTA in PBS) with Human FcR Blocking Reagent (Miltenyi 130-059-901) to block non-specific binding of antibodies. Immune and endothelial cells were overrepresented in the lung single-cell suspensions. To partially deplete these populations in the human lung samples, we stained cells isolated from lung with MACS microbeads conjugated to CD31 and CD45 (1:100, Miltenyi 130-045-801; 1:50, Miltenyi 130-091-935), washed twice with MACS buffer (300g for 5 minutes), and then passed them through an LS MACS column (Miltenyi, 130-042-401) on a MidiMACS Separator magnet (Miltenyi, 130-042-302). Cells retained on the column were designated ‘immune and endothelial enriched’. The flowthrough cells were then split, with 80% immunostained for FACS (see below) and the remaining 20% stained with EPCAM microbeads (1:50, Miltenyi 130-061-101). EPCAM stained cells were passed through another LS column. Cells retained on the column were labelled ‘epithelial enriched’, and cells that flowed through were designated ‘stromal’. For each tissue region, the ‘immune and endothelial enriched’, ‘epithelial enriched’ and ‘stromal’ fractions were mixed equally before loading onto the 10x Chromium Controller.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Lung - FACS/SS2 Sample Preparation

After negative selection against immune and endothelial cells by MACS, the remaining human lung cells were incubated with TruStain FcX (Biolegend 422302) for 5 min and stained with directly conjugated anti-human CD45-FITC (1:20, Biolegend 304038), CD31-APC (1:20, Biolegend 102410) and EpCAM-PE (1:20, Biolegend 324206) antibodies on a Nutator for 30 min at the manufacturer's recommended concentration. Cells were then pelleted (300g, 5 minutes, 4°C), washed with FACS buffer (2% FBS in PBS) three times, then incubated with cell viability marker Sytox blue (1:3,000, ThermoFisher S34857) and loaded onto a Sony SH800S cell sorter. Living single cells (Sytox blue-negative) were sorted into lysis plates based on four gates: EPCAM+CD45−CD31- (designated epithelial), EPCAM−CD45+CD31- (designated immune), EPCAM−CD45-CD31+ (designated endothelial), and EPCAM−CD45-CD31- (designated stromal).

Lymph Node – Initial tissue preparation protocol

The lymph nodes were collected in a p100 petri dish and the surrounding edges were cleaned to extract lymph nodes from the fat, using tweezers and scissors. Once all the fat was removed, the tissue was weighed. The lymph nodes were placed in a 5 mL polypropylene tube and minced with sharp scissors. Digestion media was prepared with 0.8 mg/mL Collagenase IV (Worthington) and 0.05 mg/mL DNase I (Roche) in RPMI plus 10% FBS. For every 100 mg of tissue, 1 mL of digestion media was added and transferred into a 50 mL tube, and an additional 1 mL of digestion media was added while the tissue was still being chopped. Tissue was placed in a shaker to digest at 37°C, 300 g for 20 minutes. At the end of the 20 minutes, the cell suspension was pipetted vigorously up and down 10 times to evaluate digestion and to further mechanically digest the tissue as much as possible. Additional digestion was performed when substantial undigested tissue was still present in the solution after the first round. To stop digestion, 3 µL of 0.5 M EDTA was added for every 100 mg of tissue. The tissue solution was then passed through a 100 µm cell strainer into a 15 mL tube coated with FACS buffer. If needed, a plunger was used to mash the remaining tissue in the strainer. The tube containing the tissue was washed one additional time using FACS buffer. If the cell pellet was bright red, an indication of a high concentration of red blood cells, an ACK lysis buffer step was performed (resuspend pellet in 10 mL of ACK buffer for 10 minutes at room temperature, then spin down at 4°C, 400 g for 5 minutes). After the optional ACK lysis step, cells were spun down at 4°C, 450 g for 10 minutes, resuspended in 5 mL of FACS buffer and counted.

Lymph Node – FACS/SS2 sample preparation

5 million cells were stained with CD45-FITC, Lineage-PE (CD3, CD19, CD20, CD335, CD66b) and CD31-APC. Sytox blue was added and live, single cells were sorted based on CD45-CD31+, CD45-CD31+, CD45+Lineage+ and CD45+Lineage-.

Lymph Node – 10x Genomics sample preparation

10 million cells were stained with purified mouse anti-human CD3, CD19, CD20, and CD335 for 30 minutes on ice and washed twice with media. Cells were then negatively selected by adding pan anti-mouse Dynabeads (ThermoFisher 11041) using 4 Dynabeads per target cell following manufacturer's instructions. Negatively selected cells

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

were counted and mixed with whole lymph node suspension in a 1:1 ratio. Cell concentration was adjusted to 10^6 cells/mL prior to running the sample in the 10x Genomics controller.

Mammary Glands – Initial Tissue preparation Protocol

Glands were dissected and minced with a blade into fine pieces in a 10cm petri dish. Collagenase/hyaluronidase (500µL) and 1 mL (1000 Units) DNase per 1-2g of tissue were added in 9 mL digestion media (DMEM/F12 1:1 + PSA). Tissue was digested for 3-4 hours with pipetting every 15 minutes. The digested tissue was transferred to a 50 mL Falcon tube and FACS buffer was added (2% BCS in HBSS + HEPES + PSA) to get to 30 mL of total volume. Digested tissue was spun down in a centrifuge at 1400rpm, for 5 minutes at 4°C and the supernatant was discarded. Red blood cells were lysed with ACK lysis buffer 5 mL for 5 minutes. FACS buffer was added to stop the reaction. The reaction was spun down in the centrifuge at 1400 rpm, for 5 minutes at 4°C and the supernatant was discarded. 5 mL of warmed trypsin was added to the previous reaction for 1 minute at room temp. FACS buffer was added and the tube was spun down and the supernatant was discarded. 2mL dispase and 1mL DNase were added to the mix for 2 minutes, while continuously pipetting with a 1 mL pipette. FACS buffer was added and the cells were filtered via a 40 µM mesh.

Mammary Glands – FACS/SS2 sample preparation

Cells were resuspended with FACS buffer and incubated with CD45-FITC, CD31-APC, and EPCAM-PE antibody solution for 30-45 minutes on ice. After antibody incubation, cells were spun down and washed with FACS buffer. After the washing step, cells were resuspended in FACS buffer prior to FACS sorting.

Mammary Glands – 10x Genomics Sample preparation

Cells were counted with a hemocytometer and diluted to 10^6 cells/mL and processed with 10x Genomics.

Pancreas

Exocrine pancreas – Initial tissue preparation protocol

Exocrine pancreas was put into 7 mL of digestion solution containing 2 mg/mL of collagenase VIII and 0.2 mg/mL of trypsin inhibitor in PBS. Pancreas was inflated with the digestion solution using a 27-g needle; it was then minced into small pieces and incubated in a shaker at 37°C for 8-12 minutes. After incubation, 5 mL of FACS buffer (2% FBS in PBS) was added to the tissue solution to inactivate the collagenase and cells were spun down. Cells were resuspended in 5-10 mL FACS buffer and filtered through a 100-µm cell strainer, using the back side of a syringe to further disassociate solid pieces. The cell strainer was washed with 5 mL of FACS buffer and cells were washed again, for a total of two wash steps. Cells were spun down and resuspended in 3-5 mL of 1x Red Blood Cell Lysis buffer and incubated for 8-10 minutes at room temperature. An additional 10 mL of FACS buffer were added to the cell suspension, which was then spun down.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Exocrine Pancreas – 10x Genomics sample preparation

The cells reserved for 10x were resuspended in Accutase and incubated for 10 minutes. After Accutase incubation, cells were spun down and washed twice with MACS buffer. Following the MACS manufacturer's protocol, cells were run through LS-MACS columns on a magnet, and stained with EPCAM, CD45, and CD31 single antibodies in the described order, and flow through was collected. After this first MACS step, the MACS column was removed from the magnet, samples were eluted using MACS buffer, collected, and cell density was measured using a hemocytometer. Cells that were EPCAM+, CD45+, CD31+, and triple negative cells were mixed in a 1:1:1:1 ratio. Cell concentration was adjusted to 10^6 cells/mL prior to running the sample in the 10x Genomics controller.

Exocrine Pancreas – FACS/SS2 sample preparation

Cells were resuspended with FACS buffer and incubated with CD45-FITC, CD31-APC, and EPCAM-PE antibody solution for 30-45 minutes on ice. After antibody incubation, cells were spun down and washed twice with FACS buffer. After the washing steps, cells were resuspended in FACS buffer with 10 µg/mL of DNAseI prior to FACS sorting.

Endocrine Pancreas - Initial Tissue preparation protocol

The donor pancreas was perfused with Liberase DL (1 mg/mL; Roche Diagnostics) in Hank's Buffered Salt Solution (HBSS; Gibco). The pancreas was then enzymatically digested in a water bath at 37°C until the tissue became fully dissociated. The crude pancreas digest was washed three times with wash buffer (HBSS + 0.625% human serum albumin + 10mM HEPES). At all washing steps cells were spun at 220g for 1 minute. Following the washes, the crude digest was purified on a continuous density iodixanol gradient (Alere Technologies). Cells were recovered from the gradient purification and then washed three times with wash buffer. The final washed product is enriched with pancreatic islets. The islet suspension was spun at 4°C, 425g for 3 minutes. The supernatant was carefully aspirated. Suspension was washed once with 1 mL of cold PBS, then spun at 4°C, 425g for 3 minutes and finally the supernatant was removed again. Accumax (1mL per 1000 IEQ islets) were added to resuspend the islets pellet and incubated at 37°C in a water bath to disperse islets into single cells. Gently pipet up and down to break clumps during incubation. After 3 minutes, small aliquots (3~5 µL) of suspension were taken every minute to check under a dissecting scope and evaluate the extent of dispersion. The digestion was stopped when there was not an evident increase in the proportion of single cells compared to the previous samplings. Equal volume of cold PBS was added to the final digested solution and then spun at 4°C, 1300g for 3 minutes. The supernatant was removed and a second washing step was performed. Cells were passed through a 70µm cell strainer and resuspended in 1 mL cold PBS. Cells were then ready for 10x and FACS staining prior to SS2.

Endocrine Pancreas - FACS Enrichment for SS2 and for 10x

Dissociated cells were stained with SYTOX Blue Dead Cell Stain (Thermo Fisher S34857) in Sorting Buffer (2% FBS, 10mM EGTA in PBS with 1:20 v/v RiboLock RNase inhibitor) on ice for 30 minutes in dark then washed three times. All wash steps in this section were performed using Washing Buffer (2% FBS, 10mM EGTA in PBS with 1:100 v/v RiboLock

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

RNase inhibitor) followed by centrifugation at 4°C, 1300g for 3 minutes. Cells were then resuspended in Sorting Buffer, then blocked with Human TruStain FcX (Biolegend 422302) and incubated with HPi2-Biotin for 30 minutes and washed twice. Next, cells were incubated with CD31-APC, CD45-FITC, EpCAM-PE, HPx1-AF700, and streptavidin-BV605. After washing twice, cells were sorted on a ARIA II (BD) for islet (EpCAM+HPi2), acinar (EpCAM+HPx1+), ductal (EpCAM+HPx1-HPi2-), leukocytes (CD45+), endothelial cells (CD31+) and stromal (all negative) populations. Enriched subpopulations were then submitted for SS2 sorting on SONY SH800 or loaded for 10x Genomics analysis.

Prostate - Initial Tissue preparation protocol

Digestion media (36 mg collagenese + 10 mL of HBSS with 1% HEPES) was prepared and pre-warmed at 37°C for 20 minutes. Initial tissue was placed in a petri dish on ice, rinsed 2x with PBS and a final 0.5g of tissue was prepared for dissociation. Tissue was divided in half and sections were cut from central and peripheral zones. Sections were diced for a maximum of 5 minutes to preserve tissue integrity. Sections were placed into 2 50 mL conical tubes; in each of the tubes, 10 mL of RPMI with 10% FBS was added. Tubes were spun at 12,000 rpm for 5 minutes at 4°C. Supernatant was removed and cells were resuspended in 10 mL of digestive media. Cells were then incubated for 30 minutes at 37°C with continuous end over end rotation. Final digested products were pipetted up and down 5 times to break up any clumps of tissue and subsequently passed through a 70 µm filter. Cells were washed 2x with RMPI plus 10% FBS, combined into 1 eppendorf tube, and resuspended in at least 200 µL of final volume. Cells were then ready for 10x and FACS staining prior to SS2.

Prostate - FACS/SS2 sample preparation

Cells were spun down at 1200 rpm for 5 minutes at 4°C, and the supernatant was removed. 5µL of Human TruStain FcX™ per million cells were added in 100µL of MACS buffer (PBS, 2% calf serum, 1mM EDTA). Cells were incubated on ice for 10 minutes prior to staining with antibodies of interest. Cell suspension was divided into two FACS tubes, filtering the solution using the filter on the tube cap. The following antibodies were added to each tube to a final concentration of 1:100: CD45-FITC (Biolegend 304038), EPCAM-PE (Biolegend 324206) and CD31-APC (Biolegend 303116). Cells were incubated on ice for 30 minutes. After incubation, cells were washed with 2 mL of MACS buffer, spun at 50g for 5 minutes at 4°C, the supernatant was removed and cells were resuspended in 1 mL of MACS buffer.

Salivary Gland (Parotid and Sublingual) – Initial Tissue Preparation Protocol

Salivary glands were dissociated with 1x Collagenase/Hyaluronidase in Dulbecco's Modified Eagle's Medium (DMEM) mix (Stem cell technologies, # # 07912) with 1U/mL DNase I (Stem cell technologies , #07900) for 30 minutes at 37°C. Dissociated cells were spun down and resuspended in 0.5 U/mL Dispase (Stem cell technologies, # 07913) at 37°C for 10 minutes and spun down. Cell pellet was resuspended in 1 mL of trypsin 0.05% (Gibco, #25-300-062) for 5 minutes with gentle pipetting until cell clumps were dispersed into single cells. Cell pellet was treated with 1 mL of ACK lysis buffer for 3 minutes on ice and spun down. Cells were strained through a 40 µm filter and resuspended in MACS

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

buffer (0.5% bovine serum albumin (BSA), and 2mM EDTA in PBS) and incubated with the following microbeads in the specified order for magnetic pre-enrichment as per manufacturer's protocol: 1) FcR Blocking Reagent, human (Miltenyi # 130-059-901); 2) CD45 (Miltenyi #130-045-801) and CD31 (Miltenyi # 130-091-935); 3) Epcam (130-061-101) and all triple negative flow-through from LS columns (Miltenyi #130-042-401) were collected as stromal cells. Enriched cells were washed once with 0.04% BSA in PBS and resuspended in the same buffer and counted with hemocytometer and mixed in the equal ratio for 10x Chromium capture. About 6000 cells were targeted for capture on a 10x Chromium single cell gene expression chip. For FACS sorting, cells were stained with Epcam-PE (Biolegend #324206), CD45-FITC (Biolegend #304038), CD31-APC (Biolegend #303116) and Sytox Blue (ThermoFisher #S34857) for sorting on SONY sorter for epithelial, immune, endothelial and stromal cells.

Skeletal Muscle – Initial tissue preparation protocol

The following solutions were prepared beforehand: wash medium (1 bottle of Ham's F10 medium plus 5 mL Pen/Strep solution (1%) and 50 mL Horse Serum (10%)), Dispase solution (1 g Dispase at 11 U/mL added to 100 mL of PBS and filtered with a 0.45-µm filter), Collagenase solution (500 mg of Collagenase II at 1000 U/mL added to 100 mL of PBS and filtered with a 0.45-µm filter) and Digestion medium (20 mL of wash medium with 50 mg of Collagenase II powder in a 50-mL Falcon tube). The muscle tissue was washed using the wash medium and excess fluid was removed. Tissue was then moved to a Petri dish and cut using fine scissors. The procedure was completed when no clear chunks of tissue remained visible. Curved forceps were used to scoop up the minced muscle into the 50-mL Falcon tube containing 20 mL of Digestion medium. The tissue was then incubated in a shaking water bath at 37°C for 1 h, after which the muscle sample was removed from the water bath and dried outside of the Falcon tube. Tissue was washed with 45 mL of wash medium and spun down at 1600g for 5 minutes. Cells were removed from the centrifuge and a visible pellet was identified, with some portion of fat tissue floating in the supernatant. The supernatant was removed and 16 mL of the cell suspension solution were preserved. 2 mL of Dispase Solution and 2 mL of Collagenase II solution were added to the 16 mL of cell suspension; the final solution was then vortexed to dissolve the tissue pellet. A second incubation was performed for 30 minutes in a shaking water bath at 37°C.

The samples were then removed from the water bath and the tubes were dried. The cell suspensions were divided into two 50-mL Falcon tubes, each containing 10 mL of cell suspension. Cells were further dissociated using a 10-mL syringe with a 20 gauge needle. Pieces of muscle that were too big to pass through the needle were discarded. The whole cell suspension was passed through the needle 5 times until pieces of tissue were no longer clogging the needle. The dissociated tissue was passed through a 100-µm filter and subsequently through a 40-µm filter. The filters were washed with 30 mL of wash medium, leaving the cells in a final volume of 50 mL. Cells were then spun down at 1600g for 5 minutes. After centrifugation, the cells were clearly identifiable in a pellet at the bottom of the Falcon tube. The supernatant was aspirated and the pellet was dissolved in 1 mL of wash medium.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Skin – Initial tissue protocol preparation

Collagenase IV (Worthington) stock was prepared using HBSS and the 50 mass / % volume mg/mL. DNase I (Roche) stock was prepared using PBS and 10 mass / % volume mg/mL. Skin digestion media (good for 25 g of skin) was prepared using 49 mL RPMI plus 1 mL of Collagenase IV and 150 µL of DNAse I. Digestion media was prepared no more than 10 minutes before usage. Skin was cut into pieces of 10 x 10 cm using a scalpel and a blade. Adipose subcutaneous tissue was removed using a blade scalpel swiping horizontally and without cutting the dermis. Skin was rinsed with ice cold PBS and skin was weighed. Skin was then further cut into squares (0.5 x 0.5 cm), excess PBS was removed using a metal strainer (2 mm) and transferred to a 250 mL polypropylene bottle (maximum 25 g of skin for each bottle). For every 25 g of skin, 50 mL of digestion cocktail was added and scaled down when necessary. Skin was digested for 2 hours at 37°C in a shaker at 260 rpm. A metal strainer (2 mm) was used to separate the partially digested skin tissue from the digestion media. The partially digested skin was collected in a large petri dish (150 mm) and minced with scissors. The minced skin was added back to the bottle with the digestion media. Skin was digested for an additional 8 to 10 hours (overnight) in a shaker at 37°C shaking at 260 rpm. An additional straining and 2 hours digestion was performed, with a total digestion time of less than 14 hours. Skin was not completely digested in order to have an optimal yield of leukocytes and reduced cell death. Digestion was stopped by adding EDTA to a final concentration of 5 mM and the bottle was placed on ice. Undigested skin was separated from the cell suspension using a metal strainer and subsequently passed through a 100 µm cell strainer into 50 mL tubes (3 for every 25 g of skin). The cell suspension was centrifuged at 450 g for 10 minutes at 4°C. Most of the supernatant was discarded, leaving a final volume of 10 mL. 30 mL of PBS were added to the cell suspension that was further centrifuged at 450 g for 10 minutes at 4°C. Most of the supernatant was discarded leaving only 3 mL of media behind. The pellet was resuspended with 10 mL of RPMI plus 10% FBS and filtered through a 70 mm cell strainer into a new 50 mL tube. Cells were centrifuged for 10 minutes at 450 g at 4°C. Supernatant was discarded and the pellet resuspended in 10 mL of RPMI with 10% FBS. Cells were ready for 10x and FACS staining prior to smartseq2.

Skin – FACS/SS2 sample preparation

For smartseq2, cells were stained with CD45-PE, HLA-DR-FITC, and Desmoglein-3-Alexa647. Sytox blue was added and live, single cells were sorted based on CD45+HLA-DR+, CD45+HLA-DR-, CD45-Desmoglein-3+ and CD45-Desmoglein-3-.

Skin – 10x Genomics sample preparation

Approximately 10 million total cells were stained with CD45-PE. Live single cells were sorted based on CD45+ and CD45- and combined in a 1:1 ratio. Cell concentration was adjusted to 10^6 cells/mL prior to running the sample in the 10x Genomics controller.

Spleen – Initial tissue preparation protocol

The spleen tissue was placed in a p65 petri dish and weighed. For every 3 g of tissue, 10 mL of digestion media was prepared with 0.8 mg/mL Collagenase IV (Worthington) and 0.05 mg/mL DNase I (Roche) in RPMI with 10% FBS. The spleen was then perfused by injecting digestion media with a syringe. Next, a surgical scalpel was used to slice the

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

tissue and remove cauterized sections. After slicing, the tissue was minced with scissors and transferred to a 50 mL conical tube with 5 mL digestion media. The tissue was further minced inside the 50 mL tube. The petri dish containing the tissue was washed with 5 mL of digestion media, and everything was transferred to the 50 mL tube. The tissue was digested in a shaker at 37°C, 300 g for 30 minutes. The sample was vortexed every ten minutes to re-disperse the tissue. After digestion, 100 µL of 0.5 M EDTA was added for every 10 mL of final volume, and pipetted up and down using a 25 mL pipet. Digested sample was passed through a 100 µm sterile cell strainer, and remaining tissue was mashed with a 10 mL syringe plunger. RPMI plus 10% FBS was used to wash the tissue. The cell suspension was spun at 4°C, 450 g for 10 minutes. Cells were then resuspended in 60 mL final volume and prepared for density gradient centrifugation. Ficoll-Paque PLUS (GE Healthcare) was used to remove red blood cells and decrease the amount of neutrophils, which constitute the majority of splenocytes. 12 mL of Ficoll was added to each of two 50 mL conical tubes. The splenocyte cell suspension was overlaid, and the tubes were centrifuged at 20°C, 400 g for 30 minutes, with decreased acceleration and no brakes. After centrifugation, 10 mL of media was removed from the top of the tubes, and a Pasteur pipette was used to recover the maximum amount of cells from the interphase. These cells were transferred to a 50 mL conical tube containing 10 mL of PBS, then the final volume was brought up to 45 mL with PBS. Cells were additionally centrifuged at 20°C, 200 g for 20 minutes.

Spleen – FACS/SS2 sample preparation

5 million cells were stained with CD45-FITC, Lineage-PE (CD3, CD19, CD20, and CD335), CD66b-APC-Cy7 and CD31-APC. Sytox blue was added and live, single cells were sorted based on CD45-CD31-, CD45-CD31+, CD45+Lineage+CD66b-, CD45+Lineage-CD66b+ and CD45+Lineage-CD66b-.

Spleen – 10x Genomics sample preparation

10 million cells were stained with purified mouse anti-human CD3, CD19, CD20, and CD335 for 30 minutes on ice and washed twice with media. Cells were then negatively selected by adding pan anti-mouse Dynabeads (ThermoFisher 11041) using 4 Dynabeads per target cell following manufacturer's instructions. For each donor one sample consisted of negatively selected cells and one sample consisted of total (non-enriched) cells. Cell concentration was adjusted to 10^6 cells/mL prior to running the sample in the 10x Genomics controller.

Tongue - Initial Tissue Preparation protocol

Anterior tongue biopsies were collected from tongue dorsum and posterior tongue biopsies were collected from circumvallate papilla anterior to sulcus terminalis. When multiple biopsies were available, each biopsy was digested individually. Minced tongue tissue was digested sequentially at 37°C on an orbital shaker for 1 hour in Collagenase Type IV (1250 U/ml, Worthington, LS004188), Dispase (5 U/ml, BD Biosciences, 354235) and DNase I (10 U/ml, Worthington, LS006343), followed by 30 minutes with 1x TrypLE (ThermoFisher, A1217701) and DNase I, followed by 15 minutes of 0.5% Trypsin-EDTA (ThermoFisher, 15400054) with 10 µM of Y-27632 (LC Laboratories, Y-5301). During the digestion phase, cell suspension was collected every 15 minutes and kept on ice after

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

triturating undigested tissues and fresh digest buffer was added back to the remaining tissue for further digestion. Cells were then filtered through a 40 µm strainer (Falcon 352340), pelleted at 350g for 5 minutes at 4°C, resuspended in 1x ACK lysis buffer (ThermoFisher A1049201), incubated on ice for 2 minutes, and then washed and resuspended in FACS buffer. Equal number of cells from different biopsies were combined after counting. For FACS (SmartSeq2) analysis only cells from the anterior tongue were used. Cells were blocked with FcR Blocking Reagent (Miltenyi 130-059-901), stained with CD45-FITC (Biolegend 304038), CD31-APC (Biolegend 303116), EPCAM-PE (Biolegend 324206), and Sytox Blue (ThermoFisher S34857), and sorted into four individual plate designated as Immune, Endothelial, Epithelial and Stromal Compartment. For 10x analysis cells from anterior and posterior tongue were processed independently. Cells were blocked with FcR Blocking Reagent (Miltenyi 130-059-901), stained with CD45 (Miltenyi 130-045-801) and CD31 (Miltenyi 130-091-935) to enrich for Immune/Endothelial fraction using LS columns (Miltenyi 130-042-401). Cells in the flow-through were further enriched with EPCAM (130-061-101) for Epithelial fraction, with the final triple-negative flow-through as Stromal fraction. Three fractions were combined at 1:1:1 ratio for anterior and posterior tongue, washed once with 0.04% BSA in PBS then resuspended in the same buffer at 10^6 cells/mL, and processed as two independent samples.

Trachea - Initial Tissue preparation protocol

To prepare the digest buffer for human trachea dissociation, 10000U of Collagenase Type I, 100U of DNase, and 500ul of FBS was added into 25mL of 1X Dispase in HBSS. Multiple regions of the trachea, sized approximately 1-3cm^3 each, was sampled and placed into a 12.5 mL digest buffer. The digest mix was incubated for an hour, at 37°C on a shaker set at 200 rpm. The mix was gently triturated by pipetting every 20-30 minutes. The mix was centrifuged at 500g for 5 minutes using a swinging bucket rotor. The supernatant was collected and placed on ice. The digest procedure was repeated using the remaining digest buffer, and supernatant combined. The supernatant was then filtered through a 40 µM cell strainer. The digest buffer was inactivated with 15mL of DMEM containing 10% FBS and 2X antibiotics. Cell clumps were collected by centrifugation at 500g for 5 minutes, and then supernatant aspirated and discarded. The clumps were further digested with 1X Trypsin-EDTA for 5 minutes at room temperature to obtain cells. The cells were then pelleted by centrifugation at 500g for 5 minutes, and Trypsin-EDTA aspirated and discarded. Red blood cells (rbc) were lysed by adding 3 mL of rbc lysis buffer, and incubated for 5 minutes. 10 mL of FACS buffer, which is 1X PBS containing 2% FBS, 1% PSQ and Plurionic F-68, was added to inactivate digestion and rbc lysis. Cells were again pelleted by centrifugation at 500g for 5 minutes, and re-suspended with 1mL of FACS buffer. Cell concentration and viability was measured, followed by processing using the same staining and MACS enrichment protocol as described for the bladder, for FACS and 10x respectively.

Thymus - Initial Tissue preparation protocol

Liberase digestion buffer was prepared by adding 1.25mg of liberase in 1mL of RPMI with 10% FBS). The thymus was cut in small pieces, chopped and meshed on ice. The meshed tissue was transferred into two 15 mL falcon tubes. The tissue was digested

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

adding 1.5 - 2 times its volume of liberase media for 20 to 30 minutes. Tissue was checked periodically every 5 minutes to ensure proper digestion. Digestion was stopped by adding FACS buffer (2% FBS, 1% Antibiotics (Gibco 15240-062), and 10% Pluronics (ThermoFisher 24040032) in PBS)). Cells were centrifuged at 270g at 4°C for 5 minutes and resuspended. Red blood cells were lysed by adding 3mL of rbc lysis buffer, and incubated for 5 minutes. The cell suspension was centrifuged at 500g for 5 minutes at 4°C. Cells were then stained using the anti-PE CD3 PE (Miltenyi 561803) and subsequent MAC with anti-PE microbeads (Miltenyi 130-105-6391) was performed. The MACS column was equilibrated by adding 1 mL of MACS buffer with the column on the magnet; a 15-mL Falcon collection tube labelled CD3- was placed on ice below the columns. The cell suspension was applied to the column immediately after 1 mL of MACS buffer was added, being careful to allow the cell suspension and the buffer to enter the column and collect the flow through. An additional 5 mL of MACS buffer was added and allowed to pass through the columns, collecting the flow-through. This process was repeated a second time. The tube was set aside, now containing only CD3- negative cells. The column was removed from the magnet and a new 15-mL Falcon tube labeled as CD3+ was placed. 5 mL of MACS buffer was added to the column and allowed to run through the column. The sample was plunged to recover any remaining material. Both Falcon tubes were centrifuged at 500g for 5 minutes at 4°C and re-suspended with 1mL of FACS buffer. 500 µL of the cell suspension from the CD3- and 500 µL of the CD3+ fraction were submitted for 10x analysis. The remaining fractions of CD3- and CD3+ cells were stained with anti-CD31 (APC) and anti-Epcam (APC-Cy7) or anti-CD4 (FITC) and anti-CD8 (APC), respectively, using standard protocols and submitted for smart2seq analysis.

Uterus – Initial Tissue preparation protocol

The uterus was dissected longitudinally to expose the uterine cavity and the endometrial lining was scraped with a scalpel. The myometrium was sampled from the uterine wall excluding areas adjacent to the uterine serosa. Endometrial tissue was cut into approximately 1-3 mm cubes and enzymatically dissociated with collagenase and hyaluronidase at 37°C for 1 h. Epithelial fragments were separated from stromal single cells with a 40 µm sieve, and further incubated with Accutase (MilliporeSigma, Burlington, MA) for 1 h at 37°C to dissociate into single cells. Contaminating erythrocytes were lysed by incubation for 5 minutes in RBC lysis buffer (Invitrogen). Non-viable cells were depleted from the separated epithelial and stromal fractions using EasySep Dead Cell Removal (STEMCELL Technologies Inc., Cambridge, MA) and viable cells were counted by trypan blue exclusion. Equal numbers of viable endometrial epithelial and stromal cells were combined (i.e., 1:1 endometrial epithelial:stromal cells) and the final concentration adjusted to one million viable cells/mL for single cell analysis on the 10x platform. Myometrial tissue was cut into approximately 1-3 mm cubes and digested with collagenase and hyaluronidase at 37°C for 3 h followed by Accutase for 3 h. Residual undigested tissue was removed with a 40 µm sieve. Dissociated myometrial cells from the two digestion steps were pooled without further adjustment (as they do not represent distinct tissue compartments/fractions) before depletion of non-viable cells, and adjustment of the final viable cell concentration to one million viable cells/mL for 10x single cell analysis. Endometrial and myometrial cells were submitted separately for analysis on the 10x (i.e. one tube containing only endometrial cells at a 1:1 epithelial:stromal ratio,

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and another tube containing only myometrial cells). For fluorescence activated cell sorting (FACS), dissociated endometrial and myometrial cells were pooled and processed using the following fluorochrome-conjugated antibodies (Biolegend, San Diego, CA) used at 1:20 dilution: phycoerythrin (PE) anti-SUSD2; fluorescein isothiocyanate (FITC) anti-CD45; allophycocyanin (APC) anti-CD31 (PECAM); and APC-Cy7 anti-EPCAM. Dead cells were stained with SYTOX Blue (Invitrogen). The following five populations were sorted at 1 live cell per well into 384-well plates for subsequent Smart-Seq2 analysis: SUSD2+; CD31+; CD45+; EPCAM+; and SUSD2-/CD31-/CD45-/EPCAM-.

Vasculature – Initial tissue preparation protocol

Vasculature tissue (including aorta, left and right coronary arteries, and IVC) was collected in secondary containment on ice with tissues preserved in liberase. Each sample tube containing liberase was reconstituted in 10 mL of sterile water, placed on ice, and shaken in a cold room until the whole tissue was dissolved. Elastase solution was prepared at 4 U/mL. Elastase solution and liberase solution were mixed at equal volume, and DNAseI was added in a ratio 1:100. The total volume of the vascular tissue (including aorta, coronary arteries and IVC) was 20 mL of elastase/liberase digestion solution. IVC and aorta were separated, collecting 5 mL of digestion solution including IVC and 15 mL of digestion solution including the aorta. Four 5-mL Eppendorf tubes were used to collect the aorta and coronary arteries, and 1-2 5-mL Eppendorf tubes were used for the IVC. Both aorta and IVC were cut into small pieces using scissors and scalpels. Afterward, all Eppendorf tubes were put in an incubator at 37°C for 90 minutes; samples were mixed every 5 minutes using a large pipette. During this process, 200 mL of cold buffer were prepared using PBS and 5% FBS. After digestion was completed, all cell suspensions in the 5-mL Eppendorf tubes were transferred into 15-mL Falcon tubes. Into each 15-mL Falcon tube containing cell suspension, 10 mL of cold buffer were added. Cell suspensions were filtered using a 40-µm cell strainer. Cells were centrifuged at 400g for 5 minutes (these parameters were used for all cell centrifugation steps). Supernatant was removed using vacuum and pipettes. Cell pellets were resuspended using 20 mL of cold buffer and centrifuged again, for a total of two cell washes. Erythrocytes were removed via bulk RBC lysis (Thermofisher lysis protocol A2). Cells were then resuspended in 5 mL of cold buffer, for both aorta and IVC.

Vasculature – 10x Genomics Sample preparation

Cells were counted with a hemocytometer and diluted to 10^6 cells/mL and processed with 10x Genomics.

Vasculature – FACS/SS2 Sample preparation

Cell suspensions were divided into different sub aliquots containing: unstained cells, CD45-FITC/CD31-APC, CD45-FITC/CD235a-PE, CD31-APC/CD235a-PE, and all single stained cells. The erythroid specific marker anti-ter1 was used to sort and remove remaining erythrocytes. For aorta, the total staining volume was 100 µL and each antibody was used at a final concentration of 1:50. Cells were stained for 45 minutes at

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

room temperature. Cell suspensions were centrifuged, supernatant was aspirated, and samples were resuspended in 3 mL of cold buffer and filtered using the cap of a FACS tube immediately before sorting.

10x Genomics protocol

For all organs, the 10x Genomics kits used were Chromium Next GEM Single Cell 3′ Kit v3.1 or Chromium Next GEM Single Cell 5' Kit v2. For each of these kits, the protocols provided by the manufacturer were followed.

SmartSeq2 Protocol

Plate based sequencing was performed using SmartSeq2 [35], modified for 384-well plates as described in Tabula Muris[8] and Tabula Muris Senis[9] as follows.

Lysis plate preparation

Lysis plates were created by dispensing 0.4 µL lysis buffer (0.5 U Recombinant RNase Inhibitor (Takara Bio, 2313B), 0.0625% TritonTM X-100 (Sigma, 93443-100ML), 3.125 mM dNTP mix (Thermo Fisher, R0193), 3.125 µM Oligo-dT30VN (Integrated DNA Technologies, 5′AAGCAGTGGTATCAACGCAGAGTACT30VN-3′) and 1:600,000 ERCC RNA spike-in mix (Thermo Fisher, 4456740)) into 384-well hard-shell PCR plates (Biorad HSP3901) using a Mantis liquid handler (Formulatrix). 96-well lysis plates were also prepared with a 4 µL lysis buffer. All plates were sealed with AlumaSeal CS Films (Sigma-Aldrich Z722634) and spun down (3,220g, 1 minutes) and snap-frozen on dry ice. Plates were stored at −80°C until sorting.

FACS

After dissociation, single cells from each organ and tissue were isolated into 384- or 96-well plates via FACS. Most organs were sorted into 384-well plates using three SH800S (Sony) sorters. Liver cells were sorted into 96-well plates. Limb muscle and diaphragm were sorted into 384-well plates on an Aria III (Becton Dickinson) sorter. The last column of each 384 well plate was intentionally left as blanks. For most organs, single cells were selected with forward scatter, and dead cells and common cell types were excluded with a single color channel. Combinations of fluorescent antibodies were used for most organs to enrich rare cell populations (see specific tissue method section), but some were stained only for viable cells. Color compensation was used whenever necessary. On the SH800, the highest purity setting (‘Single cell’) was used for all but the rarest cell types, for which the ‘Ultrapure’ setting was used. Sorters were calibrated using FACS buffer every day before collecting any cells, and also after every eight sorted plates. For a typical sort, 1–3 ml of pre-stained cell suspension was filtered, vortexed gently, and loaded onto the FACS machine. A small number of cells were flowed at low pressure to check cell and debris concentrations. The pressure was then adjusted, flow paused, the first destination plate unsealed and loaded, and sorting started. If a cell suspension was too concentrated, it was diluted using FACS buffer or 1x PBS. For some cell types, such as hepatocytes, 96-well plates were used because it was not possible to sort individual cells accurately into 384-well plates. Immediately after sorting, plates were sealed with a pre-labelled aluminum seal, centrifuged, and flash frozen on dry ice. On average, each 384-well plate took 5 minutes to sort.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

cDNA synthesis and library preparation

cDNA synthesis was performed using the Smart-seq2 protocol[35]. In brief, 384-well plates containing single-cell lysates were thawed on ice followed by first-strand synthesis. 0.6 µL of reaction mix (16.7 U μL^{-1} SMARTScribe Reverse Transcriptase (Takara Bio, 639538), 1.67 U μL^{-1} Recombinant RNase Inhibitor (Takara Bio, 2313B), 1.67X First-Strand Buffer (Takara Bio, 639538), 1.67 µM TSO (Exiqon, 5′-AAGCAGTGGTATCAACGCAGAGTGAATrGrGrG-3′), 8.33 mM dithiothreitol (Bioworld, 40420001-1), 1.67 M Betaine (Sigma, B0300-5VL) and 10 mM MgCl2 (Sigma, M1028-10x1ML)) was added to each well using a Mantis liquid handler. Reverse transcription was carried out by incubating wells on a Biorad C1000 384 thermal-cycler (Biorad) at 42°C for 90 minutes, and stopped by heating at 70°C for 5 minutes. Subsequently, 1.5 µL of PCR mix (1.67X KAPA HiFi HotStart ReadyMix (Kapa Biosystems, KK2602), 0.17 µM of IS PCR primer (IDT, 5′-AAGCAGTGGTAT CAACGCAGAGT-3′), and 0.038 U μl^{-1} Lambda Exonuclease (NEB, M0262L)) was added to each well with a Mantis liquid handler (Formulatrix), and second-strand synthesis was performed on a Biorad C1000 384 thermal-cycler (Biorad) thermal-cycler by using the following program: 1) 37°C for 30 minutes, 2) 95°C for 3 minutes, 3) 23 cycles of 98°C for 20 s, 67°C for 15 s and 72°C for 4 minutes, and 4) 72°C for 5 minutes. In brief, tagmentation was carried out on double-stranded cDNA using TAPS-PEG (2.5X), Tn5 enzyme (homebrew) and nuclease free water. Each well contained 120 µL of tagmentation mix (0.64 µL of TAPS-PEG, 0.2 µL of Tn5, 0.36 µL of H2O) and 0.4 µL of cDNA sample. Next, these plates were incubated at 55°C for 5 minutes using a Biorad C1000 384 thermal-cycler. This reaction was neutralized by adding 0.4 µL of 0.1% SDS solution. After this, Indexing PCR reactions were performed by adding 0.4 µL of 5 µM i5 indexing primer,0.4 µL of 5 µM i7 indexing primer and 1.2 µL of PCR mix (0.08 µL of KAPA non-HotStart enzyme, 0.8 µL of 5X buffer, 0.12 µL of 10 mM dNTPs and 0.2 µL of nuclease free H2O per well). PCR amplification was carried out on a Biorad C1000 384 thermal-cycler using the following program: 1) 72°C for 3 minutes, 2) 95°C for 30 s, 3) 10 cycles of 98°C for 10 s, 67°C for 30 s and 72°C for 1 minute 4) 72°C for 5 minutes and 5) 10°C hold.

Library pooling, quality control and sequencing

After library preparation, wells of each library plate were pooled using a Mosquito liquid handler (TTP Labtech). Pooling was followed by two purifications using 0.8x, followed by 0.7x AMPure beads (Fisher, A63881). When necessary, a third purification of 0.65x was performed. Library quality was assessed using capillary electrophoresis on a Fragment Analyzer (AATI). Pools of 20 plates were combined in equal molar amounts to make the pooled sequencing library. Alternatively, for split-lane XP loading on a NovaSeq S4 flowcell, 4 pools of 5 samples were made by combining libraries of similar average fragment length.

Organ and cell coverage

Our goals were to characterize the gene expression profile of 10,000 cells from each organ and detect as many cell types as possible. As explained in detail for each organ above, about ⅔ of the organs employed a MACS based enrichment strategy to balance cell types between four compartments; epithelial, endothelial, immune, and stromal. This

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ensured abundant cell types in one compartment did not mask rare cell types in another. Two 10x reactions per organ were loaded with 7,000 cells each with the goal to yield 10,000 QC-passed cells. Four 384-well Smartseq2 plates were run per organ. In most organs, one plate was used for each compartment (epithelial, endothelial, immune, and stromal), however, to capture rare cells, some organ experts allocated cells across the four plates differently. One column on each plate was left unsorted to serve as process control. We assumed a yield of 250 QC-passed cells per plate. The use of two 10x reactions enabled some flexibility to distinguish in the data the anatomical position of the sample or allowed enrichments other than epithelial, endothelial, immune, and stromal. It also served as insurance against losing an entire organ due to a clog of the 10x chip. To reveal BCR and TCR sequences, one or two 5' 10x reactions were run in addition to the two 3' 10x reactions on blood, bone marrow, lymph node, spleen, and thymus in some donors. All together, after QC and cell type annotation, 32,701 endothelial cells, 102,580 epithelial cells, 264,009 immune cells, and 81,529 stromal cells were characterized.

Sequencing

Sequencing runs for droplet libraries were loaded onto the NovaSeq S4 flow cell in sets of 16 to 20 libraries of approximately 5,000 cells per library with the goal of generating 50,000 to 75,000 reads per cell. Plate libraries were run in sets of 20 plates on Novaseq S4 flow cells to allow generating 1M reads per cell, depending on library quality. 152 10x reactions were performed, yielding 454,069 cells passing QC, and 161 smartseq2 plates were processed, yielding 27,051 cells passing QC.

Data extraction

Sequences from the NovaSeq 6000 were de-multiplexed using bcl2fastq version 2.20.0.4.22. Reads were aligned to the Gencode Reference 30 (GRCh38) genome using STAR version 2.6.1a_08-27 with parameters TK. Gene counts were produced using HTSEQ version 0.9.1 with default parameters, except 'stranded' was set to 'false', and 'mode' was set to 'intersection-nonempty'. Sequences from the microfluidic droplet platform were de-multiplexed and aligned using CellRanger version 5.0.1, available from 10x Genomics with default parameters.

Data pre-processing and cell type annotations

Gene count tables were combined with the metadata variables using the Scanpy[36] Python package version 1.7.2. We removed cells that did not have at least 200 detected genes. For FACS we removed cells with less than 5000 counts and for droplet cells with less than 2500 UMIs. Ambient RNA and barcode hopping[37] are known problems in 10x sequencing. To remove cells generated by barcode hopping, we removed all cells sharing both the cell and transcript barcode but not the same sample barcode in each sequencing run. In order to filter out reads from ambient RNA we ran DecontX[38] separately for each organ, using default parameters and with batch correction for donor and technology. After the DecontX filtering step, we re-filtered the dataset by excluding the mitochondrial encoded genes when removing cells that did not contain the minimum number of genes and/or minimum of counts/UMIs. The filtered gene-count matrix was then used for the analysis.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In the analysis step, we first integrated the multiple batches of data to generate a unified visualization of the cells using scVI[39] from scvi-tools[40] release 0.91.1. For training the variational autoencoder neural network, we used the following hyper parameters: n_latent=50, n_layers=3, dropout_rate=0.1. We allowed each gene to have its own variance parameter by setting dispersion="gene". We trained the scVI model with all available data and corrected the batch effect associated with donor and technology. We did not correct for batch effects associated with organ even though each organ is sequenced separately because of concerns of removing biological variation by over-correction. scVI generated an harmonized latent space that was then projected to a 2D space using UMAP. We then shared the harmonized data along with the reduced dimensional latent space in a h5ad format data object compatible with both Scanpy and CellxGene. CellxGene is a data visualization tool that allows users to interactively explore any scRNAseq dataset[10]. Manual annotation was performed by tissue experts using CellxGene. Each data object contained three main components: gene count data, cell-wise metadata, and gene-wise metadata. CellxGene allows the user to color cells by any cell metadata such as donor and compartment. Cells can also be colored by gene expression data. The user can also select cells based on any meta data features, or using a lasso tool. Following each organ and/or tissue manual annotation procedure, a data object containing the new annotations was generated and the annotations were regularized to follow the cell ontology[41]. Cell types missing in the current public version of the cell ontology were added to the provisional Tabula Sapiens cell ontology (**Supplementary Table 4**).

Since Tabula Sapiens was annotated by a large number of experts, quality control (QC) was performed on the manual annotations by using a newly developed automatic annotation tool PopularVote. PopularVote implements a total of 7 automatic annotation methods to compute the majority vote prediction as well as a predictability score based on the agreement of different algorithms. The annotation methods include random forest (RF)[42], support vector machine (SVM)[42], scANVI[11] , onClass[43], and *k* nearest neighbours (*k*NN) after batch-correction using single cell harmonization methods (scVI[39], BBKNN[44], Scanorama[45]). Briefly, the algorithm hides 20% of the cell type labels from a manually annotated dataset, and uses the other 80% of the labelled cells to generate labels for the first 20% of cells. The procedure is repeated 5 times to generate a predicted label for every cell in the dataset. Next the predictions are compared with the original manual labels of those cells. For cell types that are harder to classify, it is expected that the automatic annotation algorithms make mistakes more often and disagree with each other. If the annotation is consistent with the data, cross validation analysis returns high predictability scores and high accuracy. It should be noted that consistency is not the same as accuracy: if one cell type is substituted as another one, the consistency score will remain the same. The purpose of the cross validation study is to bring attention to cell types that are potentially mis-annotated. One caveat is that for cell types that are functionally distinct but have high transcriptional similarity, the cross validation consistency might be low even when the original labels are accurate. This will require manual examination to distinguish from true mis-classification errors. PopularVote was applied to all organs in Tabula Sapiens donors 1 and 2 and predictability scores were generated for all cells by running a 5-fold cross validation. For donors 3 to 15 a draft

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

automated annotation was generated using PopularVote. This was followed by manual inspection and annotation of all tissues in this set.

Histological analysis

Sample preparation

Additional tissue sampled from the vicinity of sequenced specimens were fixed in 10% buffered formalin and paraffin embedded. Hematoxylin and eosin(H&E) stained slides were generated using standard methods. The slides were examined with a light microscope by a team of expert pathologists who morphologically identified cells belonging to different compartments (epithelial, endothelial, immune, stromal and neuroglial) and provided a rough estimate of the ratios between the compartments. The error estimate was ranked based on the spatial heterogeneity of cell types within each tissue examined.

Automated quantification

We annotated the H&E images into 4 compartments (endothelial, epithelial, immune, and stromal) using a Napari[46] based GUI written in python. The input to this tool is the segmented/binary cell in 255s and background in 0. The segmented nuclei are obtained through Hover-Net[47]. Then the segmented cell and the original H&E image are passed into the app where the pathologist can choose one of the 4 different colors using a paintbrush over a swatch of one type of compartment and move on to the next one.

Microbiome analysis

Sample collection procedure

Surgical dissection of the gastrointestinal tract was performed at the Donor Network West facility (San Ramon, CA). Staples were used to close the pyloric sphincter and rectum. Additional staples were used to separate the duodenum, jejunum, ileum, ascending colon, and sigmoid colon, resulting in 5 sections. Each intestinal section was subsequently transected. In a sterile tissue culture hood, we used an inoculating loop to extract approximately 1 g of digesta, which was then aliquoted into a sterile microcentrifuge tube. This process was repeated up to 5 times for each intestinal section, with approximately 3 inches between each sampling. Samples were immediately transferred into a Coy anaerobic chamber to remove oxygen from the tube before sealing. Samples were then moved to -80°C freezer for storage.

DNA extraction and 16S rRNA sequencing.

DNA was extracted from 200 µL of a sample using a DNeasy Ultra Clean Microbial Kit (Qiagen). 16S rRNA amplicons were generated using the Earth Microbiome Project-recommended 515F/806R primer pairs using the 5PRIME HotMasterMix (Quantabio 2200410) with the following program in a thermocycler: 94°C for 3 minutes, 35 cycles of 94°C for 45 s, 50°C for 60 s, and 72°C for 90 s, followed by 72°C for 10 minutes. PCR products were cleaned, quantified, and pooled using the UltraClean 96 PCR Cleanup kit (Qiagen 12596-4) and Quant-iT dsDNA High Sensitivity Assay kit (Invitrogen Q33120). Samples were sequenced with 250- or 300-bp reads on a MiSeq (Illumina).

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

16S rRNA aAnalysis

De-multiplexed samples were subsequently processed using R[48]. Reads were processed and filtered using DADA2[49], with truncation lengths of 240 and 180 for forward and reverse reads, respectively. Pseudo-pooling was used to resolve rare variants within the samples. Taxonomy was assigned using the SILVA database[50], and phylogenetic trees were generated using phangorn[51]. Figures were generated using phyloseq[52].

Cell Cycle & Differentiation Analysis

We assembled a high-confidence list of cell cycle markers by first collecting a list of ~700 markers [53–56]. Subsequently, we took markers that appeared in at least two studies and assigned the cell cycle phase based on majority annotation of a select list of highly informative markers for each phase of the cell cycle. Using scanpy's score_genes function[4], we make a binary decision about a cell's cycling state, using G1, S, G2, and M cell cycle markers as the gene set for cycling, and conversely, G0 makers as the gene set for non-cycling. Moreover, the cell cycle index is derived based on the log10 ratio of the number of cycling to non-cycling cells for each cell type. Scanpy (v1.5.1) was used to generate PCA and UMAP representations, as well as to perform differential gene expression analysis. velocyto[57] and scVelo (v0.2.1) dynamical model[58] were used to obtain velocity trajectories and latent time estimates.

Immunohistochemical staining for Ki67 was performed using standard automated methods. The primary antibody used was MIB-1 from Dako M7240 at a 1:200 dilution). The assay was performed on Leica (Buffalo Grove, IL) Bond III instruments using ER2 antigen retrieval and Leica Polymer Refine detection.

Endothelial compartment analysis

Using counts corrected for ambient RNA, we extracted CD31+/Acta2- cells as valid cells (removing doublets with pericytes) in the endothelial compartment. Only tissues with >200 ECs across all donors were considered. Lymphatic endothelial cells were removed from the analysis. Cells were clustered in PCA space using the top 2500 highly variant genes and further clustered in UMAP space using PCs 1 to 12. Tissue-specific EC marker genes were determined using the FindAllMarkers function in Seurat (V3)[59] and setting cut-offs of 20% minutes.pct.exprs, log2 fold change >0.1, p_val_adj <0.05. To further ensure that tissue-specific marker genes were not due to contamination from highly expressed genes in neighboring tissue-resident cell types, the median scaled expression value of each marker gene was found for each cell type within a particular tissue. A tissue-specific marker gene was only called when the median scaled expression value was highest in ECs compared to all other cell types within the tissue. Genes whose expression is known to be affected by tissue dissociation were removed [60–64] (**Supplementary Table 5**). All marker genes were further manually curated to verify endothelial specificity.

Splicing analysis

We ran SICILIAN[18] on the BAM files generated by the STAR 2.7.5a with default parameters except for twopassMode = “Basic”, chimSegmentminutes = 12, chimJunctionOverhangminutes = 10 and chimOutType = “WithinBAM SoftClip Junctions”

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(https://github.com/czbiohub/nf-sicilian/). SICILIAN employs a statistical model to assign an empirical p-value to each candidate junction extracted from the BAM file for a 10x or Smart-Seq2 sample. It then computes the median of empirical p-values across all samples within a dataset for each junction. A cutoff of 0.1 was used for calling junctions based on median empirical p-value. Junctions were called separately for each individual and 10x and Smart-Seq2 technology. To calculate the "fraction including exon" for each individual cell for *MYL6*, we divided the number of reads mapping to the junctions corresponding to exon inclusion (chr12: 56,160,320 - 56,160,626 and chr12: 56,160,670-56,161,387) by the number of reads mapping to these junctions plus the junction corresponding to exon exclusion (chr12: 56,160,320 - 56,161,387), all after deduplicating UMIs. Cells were only considered if they had > 10 reads mapping to these three junctions, and cell types were only plotted if they had > 10 cells. To calculate "average splice position" for *CD47* we considered only the 3' splice site 108,047,292. We then ranked the 5' splice sites according to their distance from the 3' splice site: 108,057,477 (rank 4), 108,051,939 (rank 3),108,050,578 (rank 2), 108,049,619 (rank 1). We then calculated the average rank of the 5' splice site corresponding to this 3' splice site for each cell, after deduplicating UMIs. We considered only cells that had > 1 reads mapping to these four junctions, and cell types that had > 5 cells.

BCR and TCR analysis

T-Cell processing

TraCeR[65] version 0.5 (https://github.com/czbiohub/nf-tracer) was used to identify T-Cell clonal populations. Tracer assemble was run with --species Hsap set. We then ran tracer summarise with --species Hsap to create the final results. The following versions for TraCeR dependencies were used: bowtie2 version 2.3.0, igblast version 1.7.0, kallisto version v0.43.1, Salmon version 0.14.1, Trinity version v2.4.0, GRCh38 reference genome. Step-by-step instructions to reproduce the processing of the data are available from GitHub.

B-Cell processing

BraCeR[66] version 0.1 (https://github.com/czbiohub/nf-bracer) was used to identify B-Cell clonal populations. Bracer assemble was run with --species Hsap set. We then ran bracer summarise with --species Hsap to create the final results. The following versions for BraCeR dependencies were used: bowtie2 version 2.3.4.1, igblast version 1.4.0, blast 2.6.0, kallisto version v0.43.1, Salmon version 0.8.2, Trinity version v2.8.5, GRCh38 reference genome. Step-by-step instructions to reproduce the processing of the data are available from GitHub.

Data Merging and Immune Repertoire Analysis

BCR and TCR processing and merging were done via snakemake[67] and code is available on github (https://github.com/michael-swift/tabula_sapiens_workflow). Analysis on the snakemake outputs were performed in JupyterLab and are available on github (https://github.com/michael-swift/tsp-immune-analysis). Briefly, all assemblies from cell ranger 6.0.1, tracer, and bracer were annotated with IgBLAST 1.8.x[68] and output in airr format. The resulting file was preprocessed Scirpy[69] and merged with single-cell gene expression data. Only cells which had a matching assembly and gene expression data

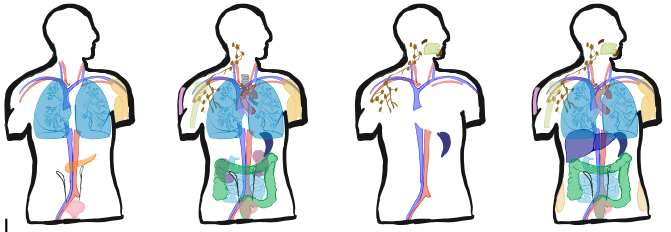
bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

were used for downstream analysis. Scirpy was used to assign clonotypes with ir.pp.ir_dist(adata, metric = “alignment”, sequence = “aa”, cutoff = 25) and ir.tl.define_clonotypes(adata, key_added=”clone_id”). Graphs of clones were created using graph tool Graph-tool[70]. A permutation test was performed to assess the empirical probability of tissue-restricted clones. Briefly, “clone_ids” were randomly shuffled 10,000 times within each donor and the number of clones present in a single tissue (tissue restricted) was counted. Isotypes were assigned to each B cell by determining the highest expressed immunoglobulin class. We summed the gene expression values in each cell for genes in each class-group (IgA, IgG, IgM/D). Subtypes were generally not resolvable due to high homology between e.g. IgA1 and IgA2. Somatic Hypermutation Frequencies were calculated as the distance from the v-gene sequence to germline v-call. Ecdf plots of mutation frequencies were created using Seaborn[71].

Tissue Immune Signature Analysis

Genes which were differentially expressed between macrophages residing in different tissues were identified using sc.tl.rank_genes_groups(adta, groupby = “tissue”, method = “wilcoxon”). Genes whose expression is known to be affected by tissue dissociation were removed [60-64] (**Supplementary Table 5**). Genes were then filtered using sc.tl.filter_rank_genes_groups(adata, minutes_fold_change = 0.8, minutes_in_group_fraction = 0.5). Genes identified in this manner were then used as the matrix for subsequent analysis, such as the correlation analysis, cluster-maps, and PCA. For plotting, the 500 most variable genes identified in the differential tissue expression analysis, sc.tl.highly_variable_genes(adata, n_top_genes = 500). The clustermap was generated using nheatmap (https://github.com/xuesoso/nheatmap).

Fig. 1



	Cell count
Bladder	21,517
Blood	50,115
Bone Marrow	12,297
Eye	10,650
Fat	20,263
Heart	11,505
Kidney	9,641
Large Intestine	13,680
Liver	5,007
Lung	35,682
Lymph Node	53,275
Mammary Gland	11,375
Muscle	30,746
Pancreas	13,497
Prostate	16,375
Salivary Gland	27,199
Skin	9,424
Small Intestine	12,467
Spleen	34,004
Thymus	33,664
Tongue	15,020
Trachea	9,522
Uterus	7,124
Vasculature	17,071

TSP1 TSP2 TSP7 TSP14 Additional 11 Donors

Number of cells
6000
12000
18000
24000
30000

Total 481,120 cells

Fig. 2

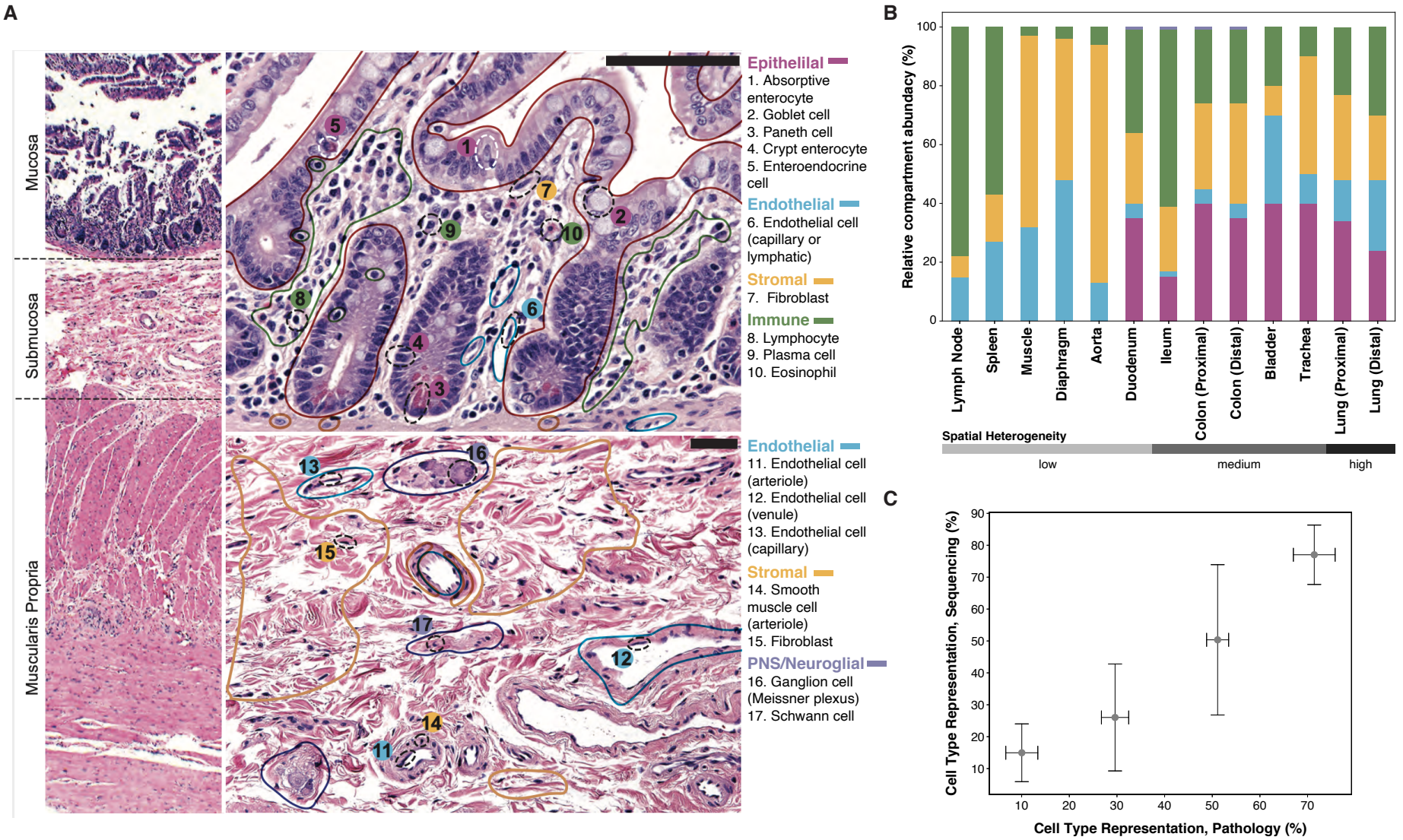

A
Mucosa
Submucosa
Muscularis Propria
Epithelilal
1. Absorptive enterocyte
2. Goblet cell
3. Paneth cell
4. Crypt enterocyte
5. Enteroendocrine cell
Endothelial
6. Endothelial cell (capillary or lymphatic)
Stromal
7. Fibroblast
Immune
8. Lymphocyte
9. Plasma cell
10. Eosinophil
Endothelial
11. Endothelial cell (arteriole)
12. Endothelial cell (venule)
13. Endothelial cell (capillary)
Stromal
14. Smooth muscle cell (arteriole)
15. Fibroblast
PNS/Neuroglial
16. Ganglion cell (Meissner plexus)
17. Schwann cell
B
Relative compartment abundacy (%)
0
20
40
60
80
100
Lymph Node
Spleen
Muscle
Diaphragm
Aorta
Duodenum
Ileum
Colon (Proximal)
Colon (Distal)
Bladder
Trachea
Lung (Proximal)
Lung (Distal)
Spatial Heterogeneity
low
medium
high
C
Cell Type Representation, Sequencing (%)
10
20
30
40
50
60
70
80
90
Cell Type Representation, Pathology (%)

Fig. 3

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

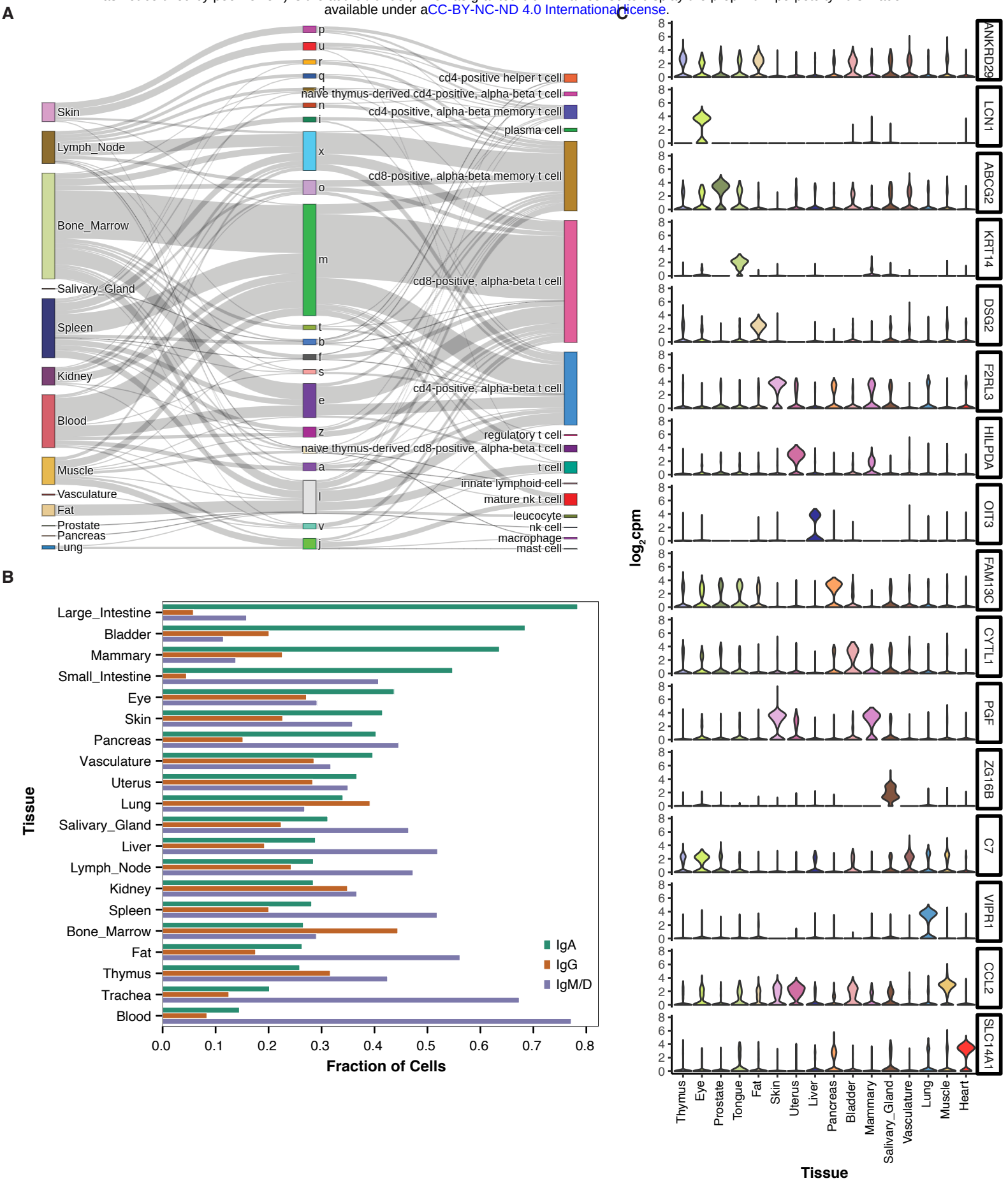


Fig. 4

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

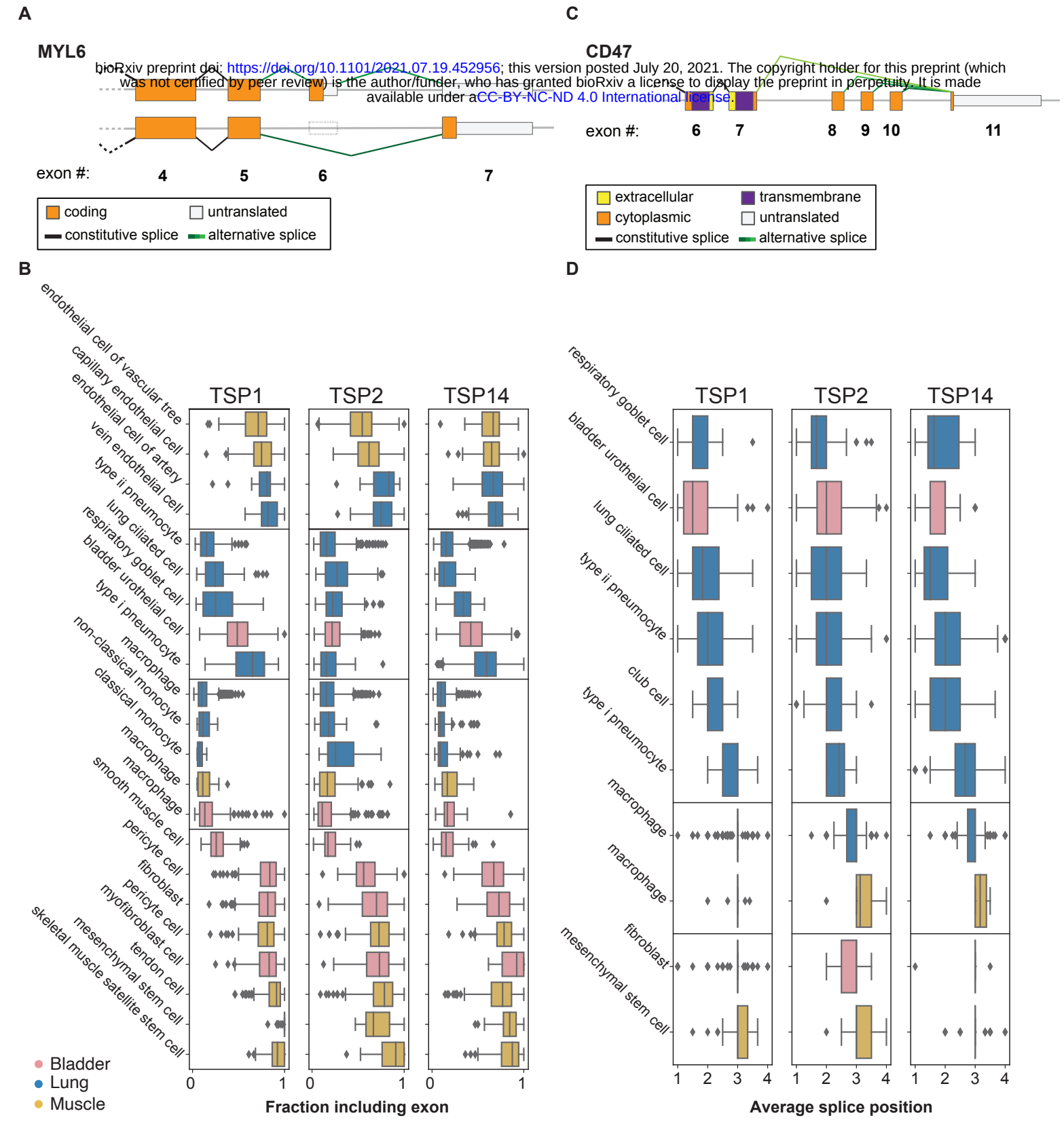


Fig. 5

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.19.452956; this version posted July 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

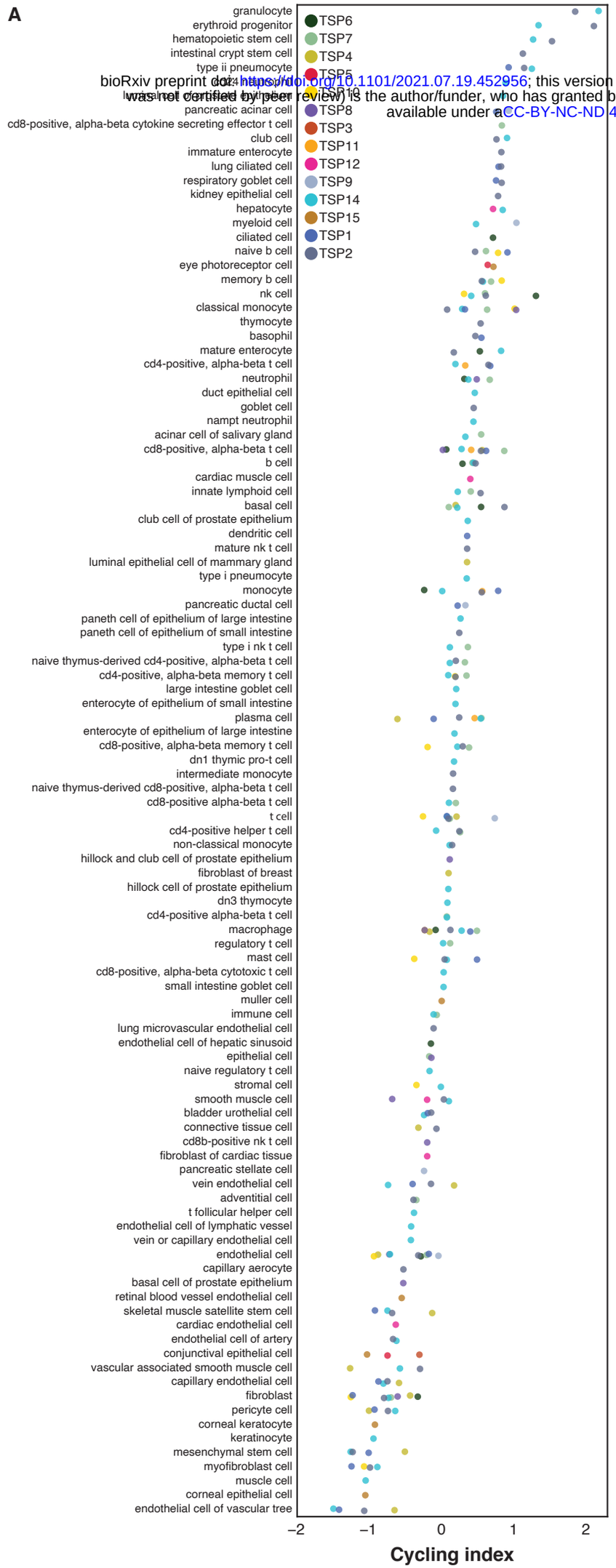
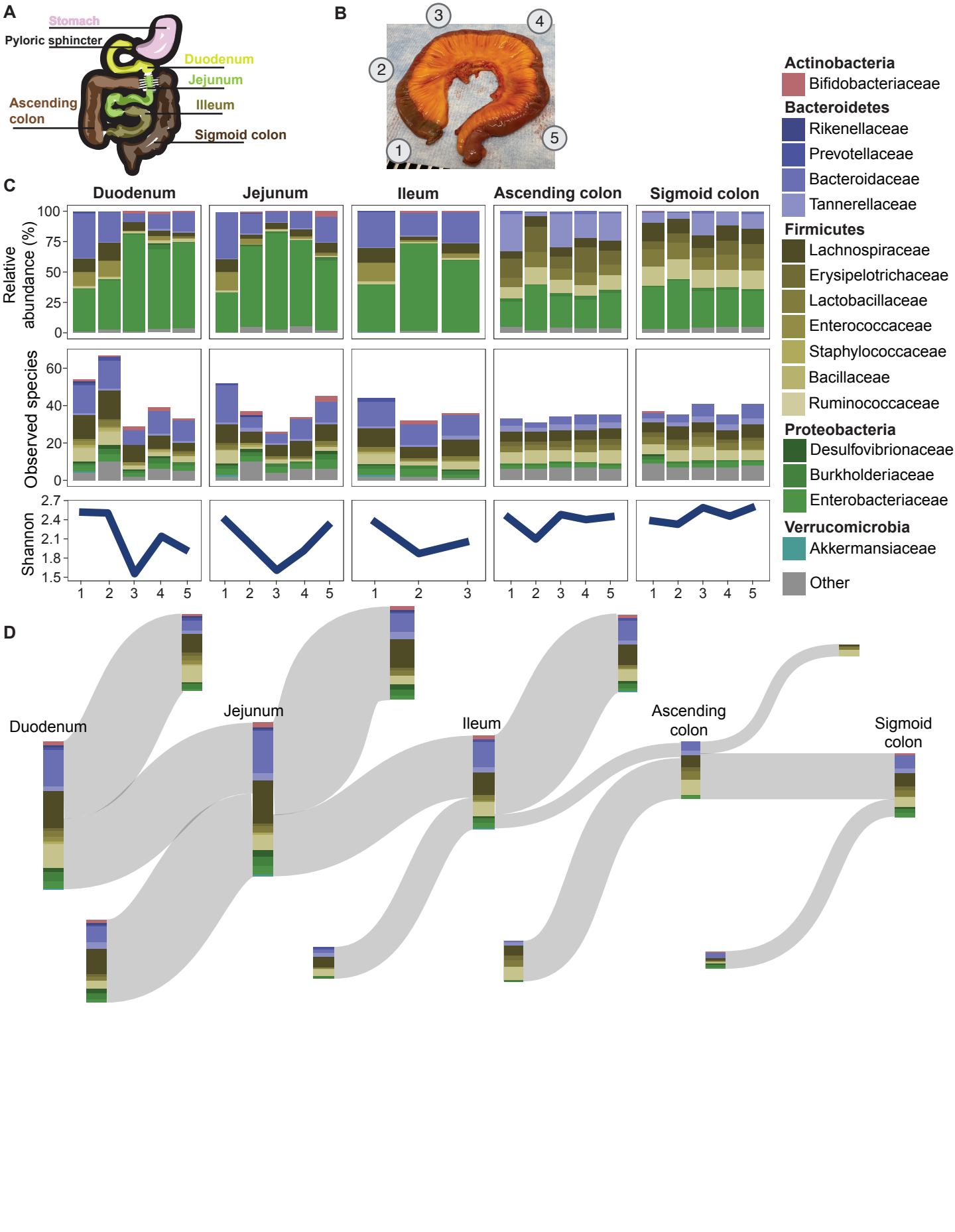


Fig. 6


A
Stomach
Pyloric sphincter
Duodenum
Jejunum
Ascending colon
Ileum
Sigmoid colon
B
1
2
3
4
5
C
Duodenum
Jejunum
Ileum
Ascending colon
Sigmoid colon
Relative abundance (%)
100
75
50
25
0
Observed species
60
40
20
0
Shannon
2.7
2.4
2.1
1.8
1.5
1 2 3 4 5
Actinobacteria
Bifidobacteriaceae
Bacteroidetes
Rikenellaceae
Prevotellaceae
Bacteroidaceae
Tannerellaceae
Firmicutes
Lachnospiraceae
Erysipelotrichaceae
Lactobacillaceae
Enterococcaceae
Staphylococcaceae
Bacillaceae
Ruminococcaceae
Proteobacteria
Desulfovibrionaceae
Burkholderiaceae
Enterobacteriaceae
Verrucomicrobia
Akkermansiaceae
Other
D
Duodenum
Jejunum
Ileum
Ascending colon
Sigmoid colon