Title

Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris

Journal

Nature, 562(7727)

ISSN

0028-0836

Authors

Schaum, Nicholas
Karkanias, Jim
Neff, Norma F
et al.

Publication Date

2018-10-01

DOI

10.1038/s41586-018-0590-4

Peer reviewed

HHS Public Access

Author manuscript

Nature. Author manuscript; available in PMC 2019 July 20.

Published in final edited form as:

Nature. 2018 October ; 562(7727): 367–372. doi:10.1038/s41586-018-0590-4.

Single-cell transcriptomics of 20 mouse organs creates a *Tabula Muris*:

The *Tabula Muris* Consortium

A full list of authors and affiliations appears at the end of the article.

Abstract

We have created a compendium of single cell transcriptomic data from the model organism *Mus musculus* comprising more than 100,000 cells from 20 organs and tissues. These data represent a new resource for cell biology, reveal gene expression in poorly characterized cell populations, and allow for direct and controlled comparison of gene expression in cell types shared between tissues, such as T-lymphocytes and endothelial cells from different anatomical locations. Two distinct technical approaches were used for most organs: one approach, microfluidic droplet-based 3'-end counting, enabled the survey of thousands of cells at relatively low coverage, while the other, FACS-based full length transcript analysis, enabled characterization of cell types with high sensitivity and coverage. The cumulative data provide the foundation for an atlas of transcriptomic cell biology.

The cell is a fundamental unit of structure and function in biology, and multicellular organisms have evolved a variety of cell types with specialized roles. Although cell types have historically been characterized by morphology and phenotype, the development of molecular methods has enabled increasingly precise descriptions of their properties, typically by measuring protein or mRNA expression patterns[1]. Technological advances have also expanded measurement multiplexing such that highly parallel sequencing can now enumerate nearly every mRNA molecule in a single cell[2–8]. This approach has provided novel insights into cell biology and organ composition from a variety of organisms[9–18]. However, while these reports provide valuable characterization of individual organs, it is challenging to compare data collected from different animals by independent labs with

*Correspondence to: quake@stanford.edu, twc@stanford.edu, spyros.darmanis@czbiohub.org.

Reviewer Information *Nature* thanks the anonymous reviewers for their contributions to the peer review of this work.

Supplementary Information is available in the online version of the paper.

Code Availability

All data, protocols, and analysis scripts from the *Tabula Muris* are shared as a public resource (http://tabula-muris.ds.czbiohub.org/). All code used for analysis is available on GitHub (https://github.com/czbiohub/tabula-muris).

Data Availability

All data, protocols, and analysis scripts from the *Tabula Muris* are shared as a public resource (http://tabula-muris.ds.czbiohub.org/). Gene counts and metadata for FACS (https://doi.org/10.6084/m9.figshare.5829687.v7) and microfluidic droplets (https://doi.org/10.6084/m9.figshare.5968960.v2) from all single cells along with all produced R objects (https://doi.org/10.6084/m9.figshare.5821263.v1), as well as FACS Index data (https://doi.org/10.6084/m9.figshare.5975392) are accessible on Figshare (https://figshare.com/projects/Tabula_Muris_Transcriptomic_characterization_of_20_organs_an d tissues from Mus musculus at single cell resolution/27733), and raw data are available on GEO (GSE109774).

varying experimental techniques. It therefore remains unknown whether these data can be synthesized as a more general resource for biology.

Here we report a compendium of cell types from the mouse *Mus musculus.* We analyzed multiple organs from the same animal, generating a dataset controlled for age, environment, and epigenetic effects. This enabled the direct comparison of cell type composition between organs, and the comparison of shared cell types across organs. The compendium is comprised of single-cell transcriptomic data from 100,605 cells isolated from 20 organs from 3 female and 4 male, C57BL/6JN, 3-month-old mice (10–15 weeks), analogous to 20-year-old humans (Fig. 1). Aorta, bladder, bone marrow, brain (cerebellum, cortex, hippocampus, striatum), diaphragm, fat (brown, gonadal, mesenteric, subcutaneous), heart, kidney, large intestine, limb muscle, liver, lung, mammary gland, pancreas, skin, spleen, thymus, tongue, and trachea from the same mouse were immediately processed into single cell suspensions. All organs were single-cell sorted into plates with FACS, and many were also loaded into microfluidic droplets (see Extended Data and Methods).

All data, protocols, analysis scripts, and an interactive data browser are publicly shared (http://tabula-muris.ds.czbiohub.org/). Gene counts and metadata are on Figshare (https://figshare.com/projects/TabulaMurisTranscriptomiccharacterizationof20organsand_tissues_from_Mus_musculus_at_single_cell_resolution/27733), raw data on GEO (GSE109774), and code is on GitHub (https://github.com/czbiohub/tabula-muris). This release allows for the exact replication of all results, facilitates in-depth analyses not completed here, and provides a comparative framework for future studies using the large variety of murine disease models. While these data are by no means a complete representation of all mouse organs and cell types, they provide a first draft attempt to create an organism-wide representation of cellular diversity.

Defining organ-specific cell types

To define cell types, we analyzed each organ independently by performing principal component analysis (PCA) on the most variable genes between cells, followed by nearest-neighbor graph-based clustering. We then used cluster-specific gene expression of known markers and genes differentially expressed between clusters to assign cell type annotations to each cluster (Extended Data Fig. 1, 2, Supplementary Table 1). All organs used a standard annotation method; an example using liver is contained in the Organ Annotation Vignette. Cell type descriptions and defining genes for each organ are available in the Supplementary Information. For each cluster, we provide annotations in the controlled vocabulary of a cell ontology[19] to facilitate inter-experiment comparisons. Many of these cell types have not previously been obtained in pure populations, and our data provide a wealth of new information on their characteristic gene expression profiles. Some unexpected discoveries include a potential new role for *Neurog3, Hex3,* and *Prss53* in the adult pancreas, a cell population expressing *Chodl* in limb muscle, transcriptional heterogeneity of brain endothelial cells, the expression of MHC class II genes by adult mouse T cells, and sets of transcription factors that distinguish cell types across organs.

Methodological comparison

We performed single-cell RNA-sequencing with two methods: FACS-based cell capture in plates and microfluidic droplet-based capture. To understand the technical biases of each approach, we performed both methods on many organs. Overall, 44,949 FACS cells and 55,656 microfluidic droplet cells were retained after quality control. Single-cell transcriptomes were sequenced to an average depth of 814,488 reads per cell (FACS) and 7,709 unique molecular identifiers (UMIs) per cell (microfluidic droplet). Comparing methods shows organ-specific differences in the number of cells analyzed (Fig. 1b, c), reads per cell (Extended Data Fig. 3a, c) and genes per cell (Extended Data Fig. 3b, d). Furthermore, with both methods the most abundant cell types analyzed are epithelial cells and leukocytes, although FACS captured a larger diversity of cell types (Extended Data Fig. 4).

Any individual single-cell sequencing experiment offers only a partial view of cell type diversity within an organism and gene expression within each cell type. We illustrate the expected variability between methods and experiments by comparing our two measurement approaches to a third method, microwell-seq, from Han *et al.*[20]. One striking feature is the variability in the number of genes detected per cell between organs and methods. For example, the median number of genes detected in bladder is ~4,900 (FACS), 2,900 (droplet), and 900 (microwell-seq), while in kidney it is ~1,400 (FACS), 1,900 (droplet), and 500 (microwell-seq). The bladder, liver, lung, mammary gland, trachea, tongue, and spleen show nearly twice as many genes detected per cell with FACS compared to microfluidic droplet, whereas heart and marrow show comparable numbers (Extended Data Fig. 5a). This difference is likely not due to sequencing depth, as both FACS and microfluidic droplet libraries are nearly saturated (Extended Data Fig. 5b). In these comparisons, a gene is considered detected if a single read maps to it, as that is the only value at which reads and UMIs can be treated equally. We also found that the number of detected genes decreases similarly across organs as the read or UMI threshold for a detectable gene is increased (Extended Data Fig. 6).

Next, we investigated whether the three methods agree on the genes defining each cell cluster (Methods). As expected, FACS and microfluidic droplet agree most, likely because they used the same biological samples. However, there are several dozen to several hundred genes common to all methods which define each cluster (Extended Data Fig. 7, Supplementary Table 2). This suggests that combining independent datasets can lead to more robust characterizations of gene expression.

Spleen and kidney are two organs for which FACS was performed without marker-based sorting, allowing us to compare the number and relative abundance of different cell types between methods. For those cell types captured by both methods, the proportion of each cell type is equivalent (Pearson correlation coefficient: spleen = 0.99, kidney = 0.99). Nonetheless, microfluidic droplets identified cell types missed by FACS in both organs, for example kidney mesangial cells, and splenic dendritic and natural killer cells. This is partially explained by cellular abundance and sampling depth (12,333 microfluidic droplet

cells vs. 2,216 FACS cells, Supplementary Table 1), and possibly from cell capture and lysis biases between methods.

As FACS captures fewer cells but detects more molecules per cell than microfluidic droplets, we asked if the two methods agree in their “bulk” gene expression profiles for the 33 shared cell populations (Methods). Such gene expression profiles largely correlate (Pearson correlation coefficient: 0.74 – 0.90), suggesting that while biases between methods exist, both accurately recapitulate average cell type gene expression profiles.

Global clustering across organs

To detect relationships between cells from different organs, we visualized all FACS cells with tSNE and grouped them with unbiased, graph-based clustering (Fig. 2, Extended Data Fig. 8). As expected, cells from different organs often mixed, with 25 of 54 clusters containing (at least five) cells from distinct organs (Fig. 3). For example, clusters 3 and 48 each contain endothelial cells from five or more organs, and clusters 1 and 24 contain mesenchymal and stromal cells from four or more organs. Cluster 2 contains B cells from fat, limb muscle, lung, spleen, marrow, and liver, but also cells annotated as leukocytes and lymphocytes from the thymus, heart, and limb muscle. This suggests that the effect of cell type on measured gene expression is stronger than the effect of batch or dissociation protocol.

Cluster co-membership alone, however, is insufficient to conclude that two cell populations from different organs represent the same or similar cell types; at any given resolution, unbiased clustering that groups related cells may also group unrelated cells[21]. Therefore, to determine which clusters are composed of related or unrelated cell types, we computed a heterogeneity score for each cluster (Methods), and found low scores for the biologically sensible clusters discussed above (Extended Data Fig. 9). In contrast, the astrocytes and epithelial cells in cluster 53 are as different from one another as two random cells.

In addition to these heterogeneous groups, the clustering reveals small populations of potentially mislabeled cells inside homogenous populations. For example, 10 thymus cells in cluster 3 (composed of 2379 cells) are annotated as ‘leukocytes’, but they are *Pecam1*$^{+}$, an endothelial marker. This is a predictable artifact of the annotation scheme: since entire clusters, rather than individual cells, were annotated in each organ, a sufficiently rare cell type that was algorithmically grouped with a more populous cell type will be mis-annotated. This seems only to occur for populations smaller than ~30 cells, which comprise under 4% of the overall dataset, and represents a lower limit of sensitivity in the current release of data interpretation.

The fact that most cells of similar cell types cluster together across organs and biological replicates shows that batch effects are not the main source of variance in the dataset. Our findings also show that manual annotation of cell types is consistent with unbiased transcriptomic clustering for sufficiently large populations. We expect that further development of multi-scale comparison algorithms will facilitate the discovery of both universal and organ-specific gene modules within these shared cell types.

To demonstrate an example of investigating common cell types across organs, we collectively analyzed all FACS cells annotated as T cells, revealing 5 clusters (Fig. 4). Cluster 0 comprises thymic cells undergoing VDJ recombination characterized by RAG (*Rag1, Rag2*) and TdT (*Dntt*) expression, and includes uncommitted double positive T-cells (*Cd4*$^{+}$*, Cd8a*$^{+}$). Cluster 4 contains predominantly proliferating thymic T cells which may represent pre-T cells expanding following VDJ recombination. Clusters 1–3 contain mostly single positive T cells (*Cd4*$^{+}$ or *Cd8a*$^{+}$). Cluster 3 contains *Cd5*^{high} thymic T cells possibly undergoing positive selection, while Cluster 2 contains mostly non-thymic T cells expressing the high affinity IL2 receptor (*Il2ra, Il2rb)*, suggesting they are activated. Interestingly, they also express MHC class II genes (*H2-Aa, H2-Ab1)*. While this is known in human T cells, MHC class II was previously thought restricted to professional antigen presenting cells in mice[22]. Finally, Cluster 1 also represents mature T cells, but primarily splenic.

Global transcription factor analysis

One major goal of defining cell identities is to understand the underlying regulatory networks. We investigated how transcription factors (TFs) contribute to cell type identity by clustering averaged gene expression profiles for each cell type using only the 1,016 TFs expressed in our dataset (Figure 5a). The resulting dendrogram closely resembles the dendrogram produced using all expressed genes, indicating that TFs can be used to reconstruct known cell ontology relationships between bulk populations (entanglement = 0.11; Extended Data Fig. 10a). In contrast, when we repeated the analysis using cell surface markers, RNA splicing factors, or the two groups combined (equivalent to a random set of genes), the entanglement was 0.22, 0.25 and 0.34, respectively, suggesting that none of these molecular classes define cell type to the extent that transcription factors do.

We then analyzed organ-specific TFs by performing correlation analysis on shared cell types between organs[23] (epithelial cells, endothelial cells, B-cells, T-cells; Fig. 5b–e, Extended Data Fig. 10b–i). To understand which TFs were most informative for specifying cell types, we performed variable selection from random forests (Methods) and determined that 136 TFs are needed to simultaneously define all cell types across all organs (Fig. 5f, Supplementary Table 3). We then determined the TF sets distinguishing each individual cell type from all other cells. These sets vary substantially in size (from 2 to 813 TFs) and are not necessarily unique to each cell type (Fig. 5g–i; Supplementary Table 4).

A possible application for such TF networks is the design of reprogramming protocols. Indeed, the TFs used in published methods are also found in the cell type-specific TFs found here (Supplementary Table 5). For some cell types, such as hepatocytes, satellite cells, and oligodendrocytes, those reprogramming factors are the top variables segregating cell types (Fig. 5g–i). In fact, for nearly all reprogramming protocols the TFs used also specified the targeted cell type in our data (Supplementary Table 5), suggesting that our data can inform novel reprogramming schemes.

Discussion

A key challenge for single-cell studies is to understand transcriptomic changes caused by dissociation. A previous study showed that quiescent limb muscle satellite cells activate upon dissociation and consequently express immediate early genes and other dissociation-related markers[24]. We clearly observed these markers in several organs including limb muscle (Extended Data Fig. 11), but many showed little evidence of cellular activation. Therefore, the dissociation- related satellite cell markers are not universal, and organs likely display unique dissociation-related expression profiles. Importantly, the presence of such gene expression changes does not prevent cell type identification or the comparison of cell types across organs.

Another challenge faced by single-cell studies is experimental design amid the choice of multiple technologies. Droplet-based technologies offer certain advantages for discovering rare cell types or states, for example when many cells (10,000s) are required to reconstruct whole-organism architecture and developmental lineages[25,26]. FACS-based methods generate high coverage over small cellular populations (10s - 1,000s), and are beneficial for enriching specific or rare cell types, and for studying subtle heterogeneity involving lowly expressed genes[27], alternative splicing[15], and sequence variation analysis[28]. There are opportunities to combine the two methods, such as by running sorted cells on a microfluidic droplet platform, which could potentially accommodate both cell type enrichment and cost.

Recently, Han et al.[20] published a complementary scRNA-seq study across mouse organs. Their data contains four times as many cells and included several sample types not present in our data, such as neonatal and fetal organs, cell lines, and young adult ovary, peripheral blood, placenta, prostate, small intestine, stomach, testis, and uterus. However, our FACS data contains four times as many genes per cell, and we analyzed several organs not present in the Han et al. data set, such as aorta, four brain regions, diaphragm, four fat types, four adult heart chambers, adult skin, tongue, and trachea. Additionally, several features of our study facilitate replication and cross-experiment analysis: 1) All data, analysis, and code is freely available, 2) Our web portal allows one to query gene expression in all organs simultaneously, 3) We annotated cell types using standard cell ontology terms, thereby enabling cross-organ and cross-experiment analyses, 4) Age and sex is controlled in our data by collecting all organs from the same animals, 5) Both sexes are represented for all organs in our data, 6) Organs were perfused allowing for the analysis of tissue- resident immune cells, and 7) Full-length transcript data makes possible transcription factor, splice variant, and sequence variant analyses.

In conclusion, we have created a compendium of single-cell transcriptional measurements across 20 mouse organs. This *Tabula Muris*, or “Mouse Atlas”, has many uses, including the discovery of new putative cell types, the discovery of novel gene expression in known cell types, and the ability to compare cell types across organs. It will also serve as a reference of healthy young adult organs which can be used as a baseline for current and future mouse models of disease. While it is not an exhaustive characterization of all mouse organs, it does provide a rich dataset of the most highly studied organs in biology. *Tabula Muris* provides a framework and description of many of the most populous and important cell populations

within the mouse, and represents a foundation for future studies across a multitude of diverse physiological disciplines.

Methods

Mice and Organ Collection

Four 10–15 week old male and four virgin female C57BL/6JN (derived from C57BL/6J in 2016) mice were shipped from the National Institute on Aging colony at Charles River to the Veterinary Medical Unit (VMU) at the VA Palo Alto (VA). At both locations, mice were housed on a 12-h light/dark cycle, and provided food and water *ad libitum.* The diet at Charles River was NIH-31, and Teklad 2918 at the VA VMU. Littermates were not recorded or tracked, and mice were housed at the VA VMU for no longer than 2 weeks before euthanasia. Prior to tissue collection, mice were placed in sterile collection chambers for 15 minutes to collect fresh fecal pellets. Following anesthetization with 2.5% v/v Avertin, mice were weighed, shaved, and blood drawn via cardiac puncture before transcardial perfusion with 20 ml PBS. Mesenteric adipose tissue (MAT) was then immediately collected to avoid exposure to the liver and pancreas perfusate, which negatively impacts cell sorting. Isolating viable single cells from both pancreas and liver of the same mouse was not possible, therefore, 2 males and 2 females were used for each. Whole organs were then dissected in the following order: large intestine, spleen, thymus, trachea, tongue, brain, heart, lung, kidney, gonadal adipose tissue (GAT), bladder, diaphragm, limb muscle (*tibialis anterior)*, skin (dorsal), subcutaneous adipose tissue (SCAT, inguinal pad), mammary glands (fat pads 2, 3, and 4), brown adipose tissue (BAT, interscapular pad), aorta, and bone marrow (spine and limb bones). Following single cell dissociation as described below, cell suspensions were either used for FACS of individual cells into 384-well plates, or for microfluidic droplet library preparation. All animal care and procedures were carried out in accordance with institutional guidelines approved by the VA Palo Alto Committee on Animal Research.

Tissue dissociation and sample preparation

Specific protocols for each tissue are described in the supplement.

Sample size, randomization, and blinding

No sample size choice was performed prior to the study. Randomization and blinding were not performed – the authors were aware of all data and metadata related variables during the entire course of the study.

Single Cell Methods

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% Triton™ X-100 (Sigma, 93443–100ML), 3.125 mM dNTP mix (Thermo Fisher, R0193), 3.125 µM Oligo-dT_{30}VN (IDT, 5’AAGCAGTGGTATCAACGCAGAGTACT$_{30}$VN-3’) and 1:600,000 ERCC RNA spike-in mix (Thermo Fisher, 4456740)) into 384-well hard-shell PCR plates (Biorad HSP3901)

using a Tempest liquid handler (Formulatrix). 96-well lysis plates were also prepared with 4 μl lysis buffer. All plates were sealed with AlumaSeal CS Films (Sigma-Aldrich Z722634) and spun down (3,220 × g, 1 minute) 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 Fluorescence Activated Cell Sorting (FACS). Most organs were sorted into 384-well plates using SH800S (Sony) sorters. Heart and liver were sorted into 96-well plates and cardiomyocytes were hand-picked into 96-well plates. Limb muscle and diaphragm were sorted into 384-well plates on an Aria III (Becton Dickinson) sorter. The last two columns of each 384 well plate were 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 for rare cell populations (see supplemental text), 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 8 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 like 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-labeled aluminum seal, centrifuged, and flash frozen on dry ice. On average, each 384-well plate took 8 minutes to sort.

cDNA synthesis and library preparation

cDNA synthesis was performed using the Smart-seq2 protocol[7,8]. Briefly, 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 SMARTScribe Reverse Transcriptase (Takara Bio, 639538), 1.67 U/μl 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 DTT (Bioworld, 40420001–1), 1.67 M Betaine (Sigma, B0300–5VL), and 10 mM $MgCl_2$ (Sigma, M1028–10X1ML)) was added to each well using a Tempest liquid handler. Reverse transcription was carried out by incubating wells on a ProFlex 2 × 384 thermal-cycler (Thermo Fisher) 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 IS PCR primer (IDT, 5’-AAGCAGTGGTATCAACGCAGAGT-3’), and 0. 038 U/μl Lambda Exonuclease (NEB, M0262L)) was added to each well with a Mantis liquid handler (Formulatrix), and second strand synthesis was performed on a ProFlex 2×384 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 seconds, 67°C for 15 seconds, and 72°C for 4 minutes, and 4) 72°C for 5 minutes.

The amplified product was diluted with a ratio of 1 part cDNA to 10 parts 10mM Tris-HCl (Thermo Fisher, 15568025), and concentrations were measured with a dye-fluorescence assay (Quant-iT dsDNA High Sensitivity kit; Thermo Fisher, Q33120) on a SpectraMax i3x microplate reader (Molecular Devices). Sample plates were selected for downstream processing if the mean concentration of blanks (ERCC-containing, non-cell wells) was greater than 0 ng/µl, and, after linear regression of the values obtained from the Quant-iT dsDNA standard curve, the R^2 value was greater than 0.98. Sample wells were then selected if their cDNA concentrations were at least one standard deviation greater than the mean concentration of the blanks. These wells were reformatted to a new 384-well plate at a concentration of 0.3 ng/µl and final volume of 0.4 µl using an Echo 550 acoustic liquid dispenser (Labcyte).

Illumina sequencing libraries were prepared as described in Darmanis et al. 2015[14]. Briefly, tagmentation was carried out on double-stranded cDNA using the Nextera XT Library Sample Preparation kit (Illumina, FC-131–1096). Each well was mixed with 0.8 µl Nextera tagmentation DNA buffer (Illumina) and 0.4 µl Tn5 enzyme (Illumina), then incubated at 55°C for 10 minutes. The reaction was stopped by adding 0.4 µl "Neutralize Tagment Buffer" (Illumina) and centrifuging at room temperature at 3,220 × g for 5 minutes. 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 Nextera NPM mix (Illumina). PCR amplification was carried out on a ProFlex 2×384 thermal cycler using the following program: 1) 72°C for 3 minutes, 2) 95°C for 30 seconds, 3) 12 cycles of 95°C for 10 seconds, 55°C for 30 seconds, and 72°C for 1 minute, and 4) 72°C for 5 minutes.

Library pooling, quality control, and sequencing

Following library preparation, wells of each library plate were pooled using a Mosquito liquid handler (TTP Labtech). Pooling was followed by two purifications using 0.7x AMPure beads (Fisher, A63881). Library quality was assessed using capillary electrophoresis on a Fragment Analyzer (AATI), and libraries were quantified by qPCR (Kapa Biosystems, KK4923) on a CFX96 Touch Real-Time PCR Detection System (Biorad). Plate pools were normalized to 2 nM and equal volumes from 10 or 20 plates were mixed together to make the sequencing sample pool. A PhiX control library was spiked in at 0.2% before sequencing.

Sequencing libraries from 384-well and 96-well plates

Libraries were sequenced on the NovaSeq 6000 Sequencing System (Illumina) using 2 × 100bp pairedend reads and 2 × 8bp or 2 × 12bp index reads with either a 200- or 300-cycle kit (Illumina, 20012861 or 20012860).

Microfluidic droplet single cell analysis

Single cells were captured in droplet emulsions using the GemCode Single-Cell Instrument (10x Genomics, Pleasanton, CA, USA), and scRNA-seq libraries were constructed as per the

10x Genomics protocol using GemCode Single-Cell 3′ Gel Bead and Library V2 Kit. Briefly, single cell suspensions were examined using an inverted microscope, and if sample quality was deemed satisfactory, the sample was diluted in PBS with 2% FBS to a concentration of 1000 cells/μl. If cell suspensions contained cell aggregates or debris, two additional washes in PBS with 2% FBS at 300 × g for 5 minutes at 4°C were performed. Cell concentration was measured either with a Moxi GO II (Orflo Technologies) or a hemocytometer. Cells were loaded in each channel with a target output of 5,000 cells per sample. All reactions were performed in the Biorad C1000 Touch Thermal cycler with 96-Deep Well Reaction Module. 12 cycles were used for cDNA amplification and sample index PCR. Amplified cDNA and final libraries were evaluated on a Fragment Analyzer using a High Sensitivity NGS Analysis Kit (Advanced Analytical). The average fragment length of 10x cDNA libraries was quantitated on a Fragment Analyzer (AATI), and by qPCR with the Kapa Library Quantification kit for Illumina. Each library was diluted to 2 nM, and equal volumes of 16 libraries were pooled for each NovaSeq sequencing run. Pools were sequenced with 100 cycle run kits with 26 bases for Read 1, 8 bases for Index 1, and 90 bases for Read 2 (Illumina 20012862). A PhiX control library was spiked in at 0.2 to 1%. Libraries were sequenced on the NovaSeq 6000 Sequencing System (Illumina)

Data Processing

Sequences from the NovaSeq were de-multiplexed using bcl2fastq version 2.19.0.316. Reads were aligned using to the mm10plus genome using STAR version 2.5.2b with parameters TK. Gene counts were produced using HTSEQ version 0.6.1p1 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 2.0.1, available from 10x Genomics with default parameters.

Clustering

Standard procedures for filtering, variable gene selection, dimensionality reduction, and clustering were performed using the Seurat package version 2.2.1. A detailed worked example, including the mathematical formulae for each operation, is in the Organ Annotation Vignette. The parameters that were tuned on a per-tissue basis (resolution and number of PCs) can be viewed in the tissue-specific Rmd files available on GitHub. For each tissue and each sequencing method (FACS and microfluidic droplet), the following steps were performed:

1. Cells were lexicographically sorted by cell ID to ensure reproducibility.
2. Cells with fewer than 500 detected genes were excluded. (A gene counts as detected if it has at least one read mapping to it). Cells with fewer than 50,000 reads (FACS) or 1000 UMI (microfluidic droplet) were excluded.
3. Counts were log-normalized for each cell using the natural logarithm of 1 + counts per million (for FACS) or 1 + counts per ten thousand (for microfluidic droplet).

4. Variable genes were selected using a threshold (0.5) for the standardized log dispersion, where the standardization was done in separately according to binned values of log mean expression.

5. The variable genes were projected onto a low-dimensional subspace using principal component analysis. The number of principal components was selected based on inspection of the plot of variance explained.

6. A shared-nearest-neighbors graph was constructed based on the Euclidean distance in the lowdimensional subspace spanned by the top principal components. Cells were clustered using a variant of the Louvain method that includes a resolution parameter in the modularity function[13].

7. Cells were visualized using a 2-dimensional t-distributed Stochastic Neighbor Embedding of the PC-projected data.

8. Cell types were assigned to each cluster using the abundance of known marker genes. Plots showing the expression of the markers for each tissue appear in the extended data.

9. When clusters appeared to be mixtures of cell types, they were refined either by increasing the resolution parameter for clustering or subsetting the data and rerunning steps 3–7.

A similar analysis was done globally for all FACS processed cells and for all microfluidic droplet processed cells to produce an unbiased clustering.

Heterogeneity Score

Let C be a cluster, decomposed into annotated cell types $C = T_1 \cup \cdots \cup T_k$. For each pair of cell types Ti,Tj, we compute the average distance between their members: $d_{ij} = \frac{1}{|T_i||T_j|} \sum_{x \in T_i, y \in T_j} x - y|$. The heterogeneity score C is the maximum of those distances over cell types T with at least five cells. For the FACS data, the vector x for a cell is the PC-projection from step 5 above. Extended Data Figure 9 contains heatmaps of the cell-type distance matrix dij for select clusters and a barplot of the heterogeneity scores for all clusters containing multiple cell types.

Differential expression overlap analysis

For FACS and microfluidic droplet data, differential expression analysis for each organ was performed using a Wilcoxon rank-sum test as implemented in the “FindAllMarkers” function of the Seurat package. Differential expression was performed between cell ontology groups and resulted in a list of differentially expressed genes (logeFoldChange > 0.25) between each cell ontology group and all other ontology groups of the same organ. For microwell-Seq we used the corresponding published lists for each cell type and for every organ. We then assessed the overlap of those lists between the three methods. As the nomenclature is not identical, the analysis was performed between cell types that could be matched with a certain degree of confidence between the three methods (Supplementary Table 2).

Correlating bulk gene expression profiles

For the 33 cell populations shared between FACS and microfluidic droplets, the average gene expression profile of each population was calculated. The quality of such a bulk gene expression profile depends on the total number of detected molecules. FACS detects more molecules per cell, but fewer cells. Microfluidic droplets detect fewer molecules per cell, but more cells. To assess the agreement between methods on annotated cell types, Pearson correlation was used on the log expression profiles of each shared cell population. (Only genes with present at 1 CPM or greater in at least one of the datasets were considered. A pseudocount of 1 CPM was added before taking logarithms.)

Calculation of dissociation scores

For each organ, principal component analysis was performed on a subset of 140 dissociation-related genes[23]. The first principal component was used as the "dissociation score" as it corresponds to the variance within these genes.

Defining cell type-enriched transcription factors

Transcription factors were defined as the 1140 genes annotated by the Gene Ontology term "DNA binding transcription factor activity", downloading from the Mouse Genome Informatics database (http://www.informatics.jax.org/mgihome/GO/project.shtml, accessed on 2017–11-10). Cell types were defined as unique combinations of cell ontology and organ annotation (e.g. Lung_Endothelial_cell). All analyses were performed on the full dataset, subsampled by randomly selecting 60 cells from each cell type. Enriched TFs were defined by the Seurat FindMarkers function with the "Wilcoxon" significance test for the target cell type against the all of rest of the cell types combined. These were filtered by p_val < 10–3, avg_diff > 0.2, pct.1 – pct.2 > 0.1 (percent detected difference > 0.1), and pct.1 > 0.3 (detected in > 30% of target cells).

Cell type comparisons between methods using cell ontology classes

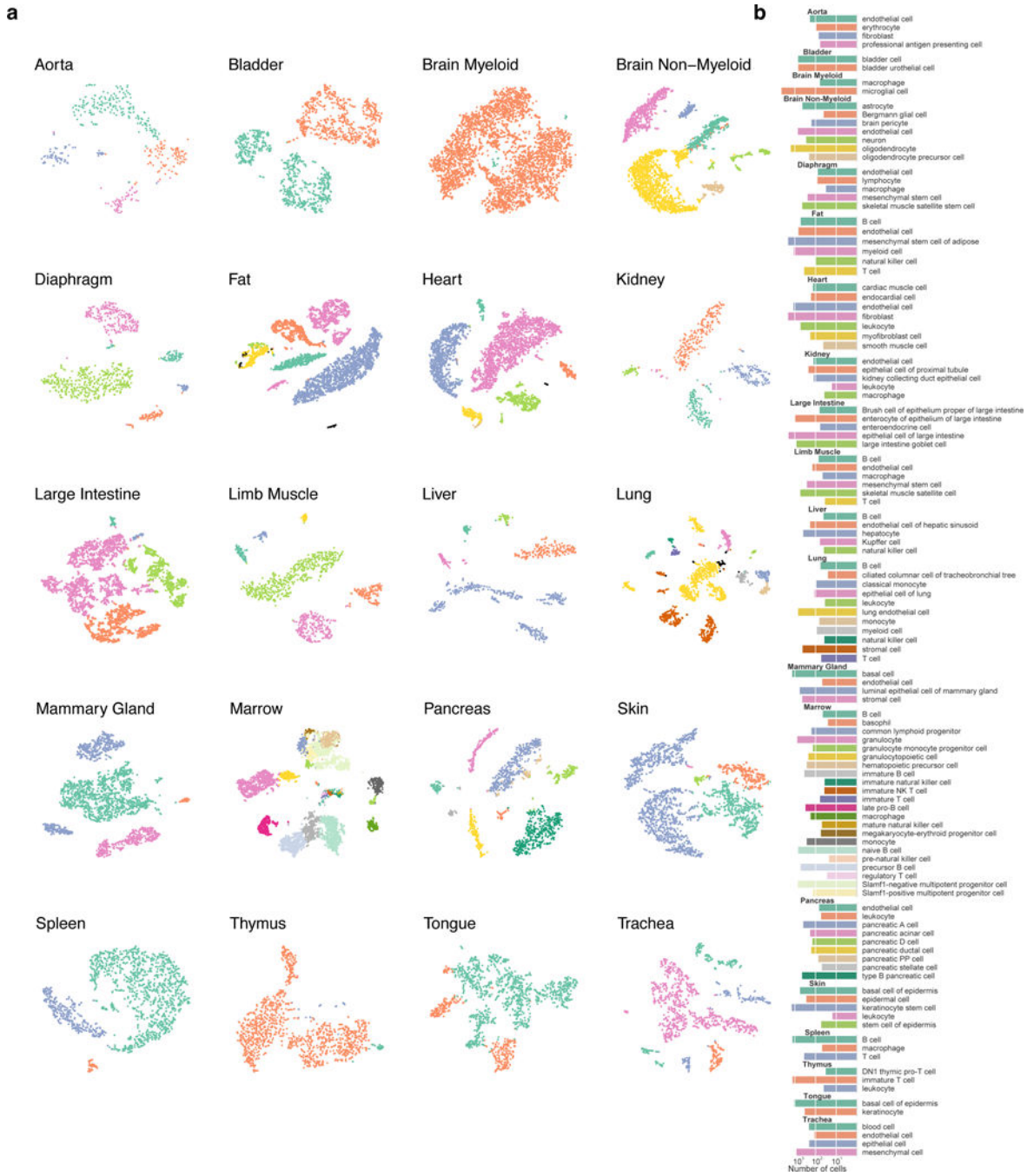
We used the OntologyX R package family version 2.4 (libraries ontologylndex, ontologyPlot, and ontologySimilarity) to draw the representative cell ontology dendrograms (function onto_plot). To compute the tanglegram (function tanglegram from dendextend R package version 1.8) we used the dendrogram created from all expressed genes as the reference for comparisons to the dendrograms produced using particular GO cellular functions (transcription factors, cell surface markers, RNA splicing factors). The entanglement scores were calculated using the step2side method (function untangle from dendextend R package). Entanglement is a measure of alignment between two dendrograms. The entanglement score ranges from 0 (exact alignment) to 1 (no alignment)[29].

Defining TF networks with random forests

We used random forests (a classifier that combines many single decision trees) to calculate the importance of each gene for defining cell types[30]. The varSelRF R package version 0.7–8 uses the out-of-bag (OOB) error as the minimization criterion and carries out variable elimination with random forests by successively eliminating the least important variables (with importance as returned from the random forest analysis). The algorithm iteratively fits

random forests, at each iteration building a new forest after discarding those variables (genes) with the smallest variable importance; the selected set of genes is the one that yields the smallest OOB error rate. This leads to the selection of small sets of non-redundant variables.

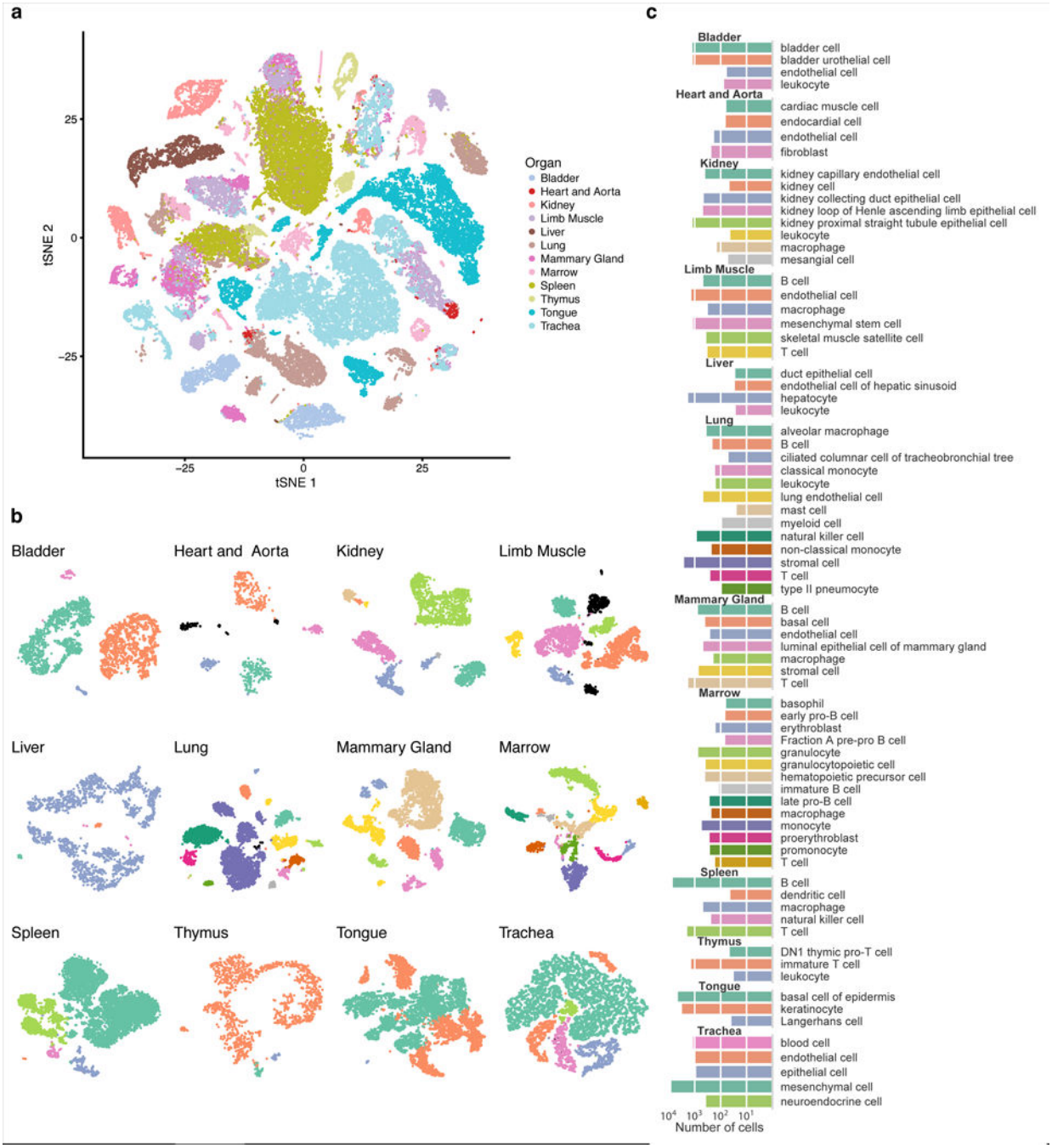
Extended Data



Extended Data Figure 1.

The number and type of FACS cells composing each organ.

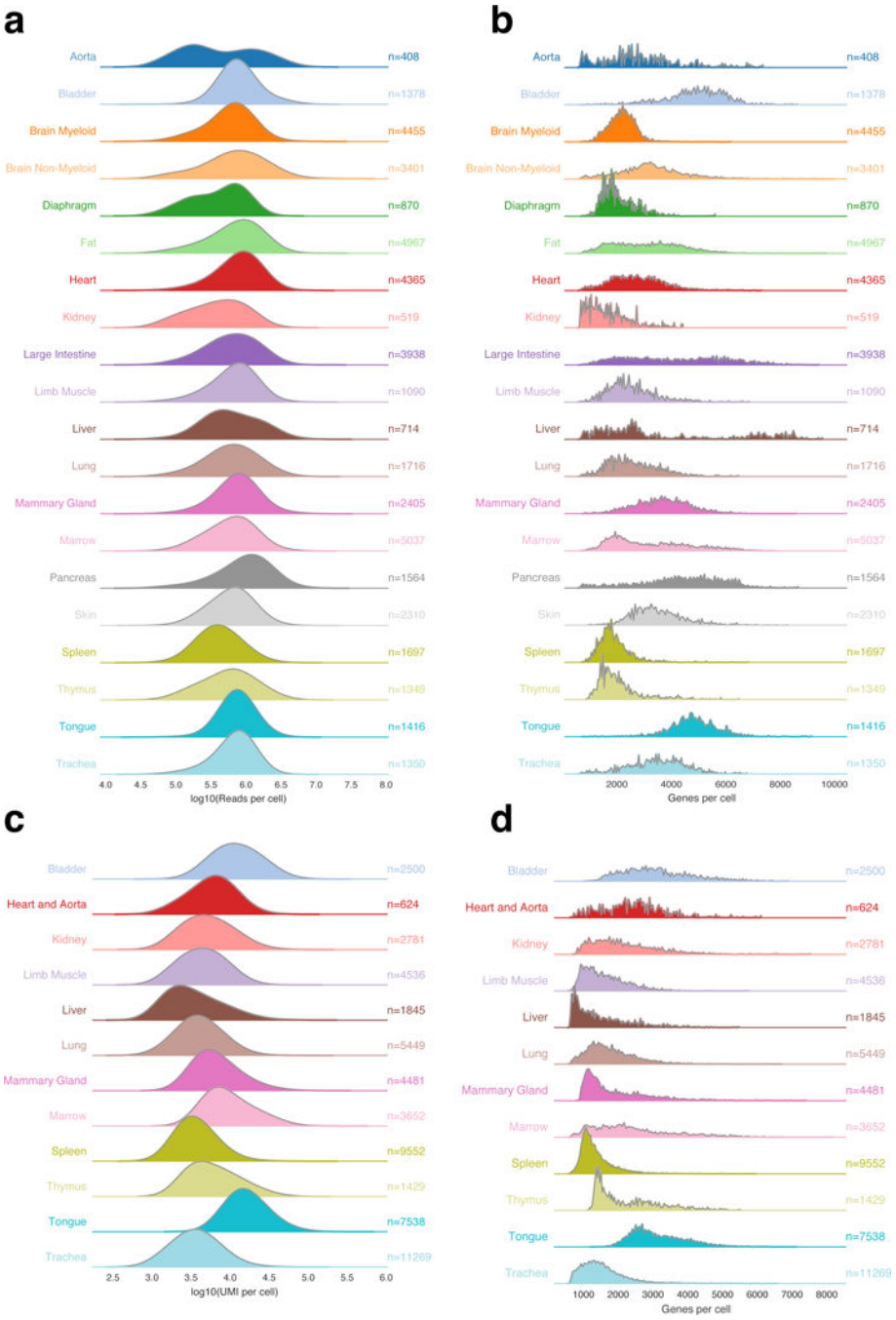
a) Cells for each organ visualized with tSNE, colored by cell type. Cell types were determined by differential gene expression of known markers between clusters. b) Barplots quantifying the number of each annotated cell type. Cell type colors match their respective tSNE plot.



Extended Data Figure 2.

The number and type of microfluidic cells composing each organ.

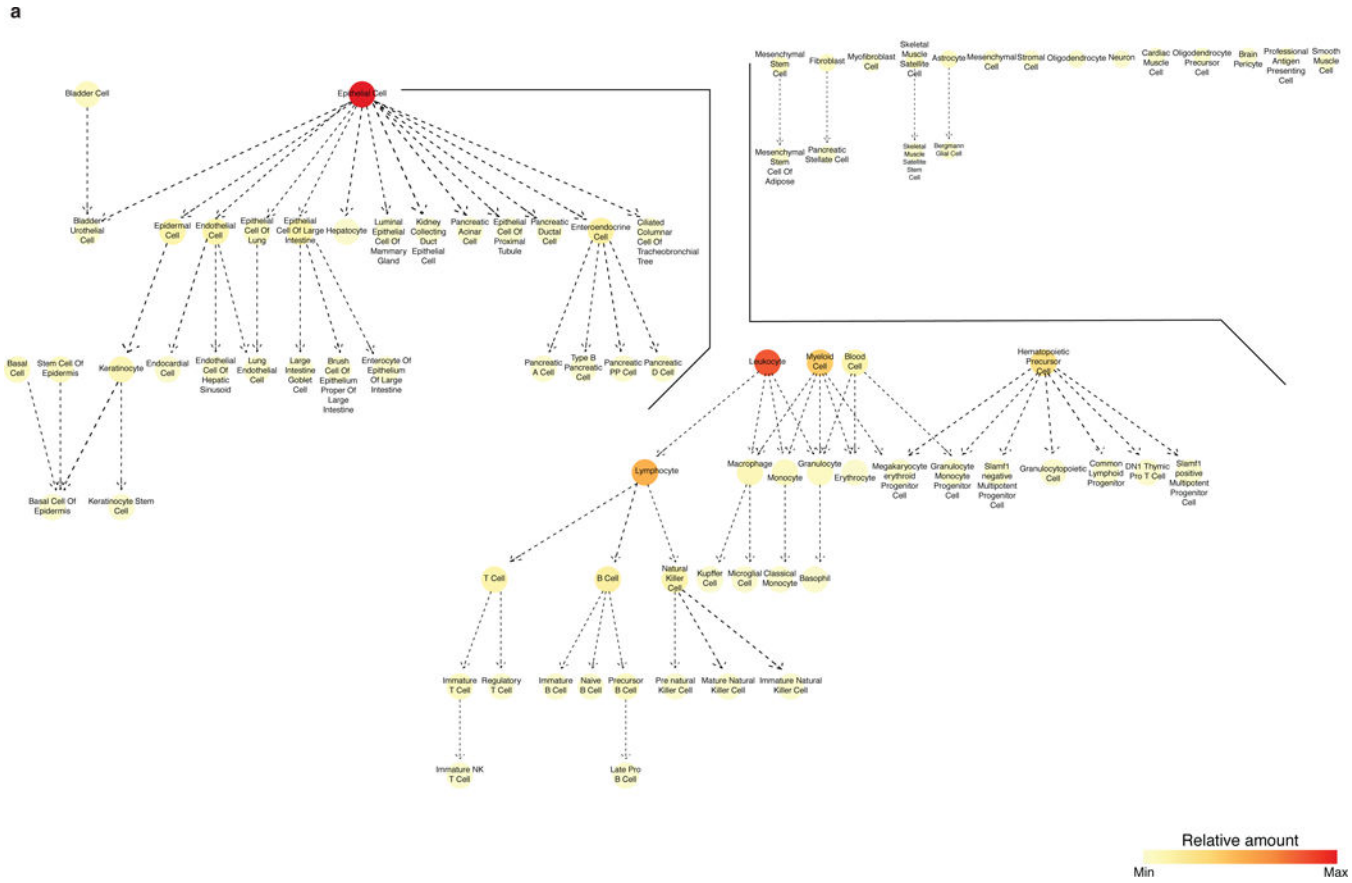
a) tSNE plot of all cells collected by microfluidic droplet, colored by organ, overlaid with the predominant cell type composing each cluster. b) Cells for each organ visualized with tSNE, colored by cell type. Cell types were determined by differential gene expression of known markers between clusters. c) Barplots quantifying the number of each annotated cell type. Cell type colors match their respective tSNE plot.

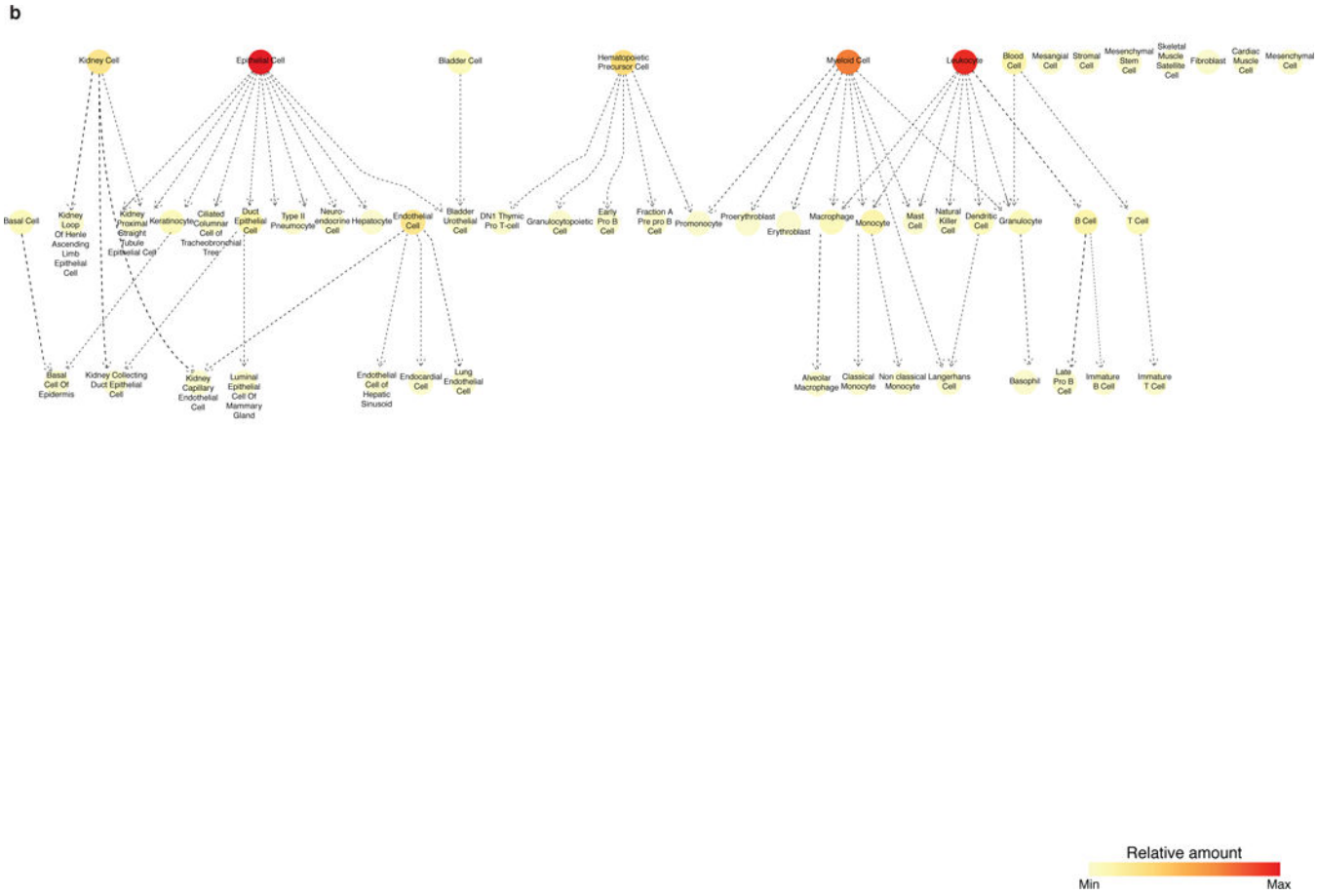


Extended Data Figure 3.

The number of reads, UMIs, and genes detected per cell for each organ.

Histogram for each organ of the number of a) reads per cell (FACS), and c) unique molecular identifiers (UMIs) per cell (microfluidic droplet). Histogram of the number of genes detected per cell for each organ from b) FACS, and d) microfluidic droplet.

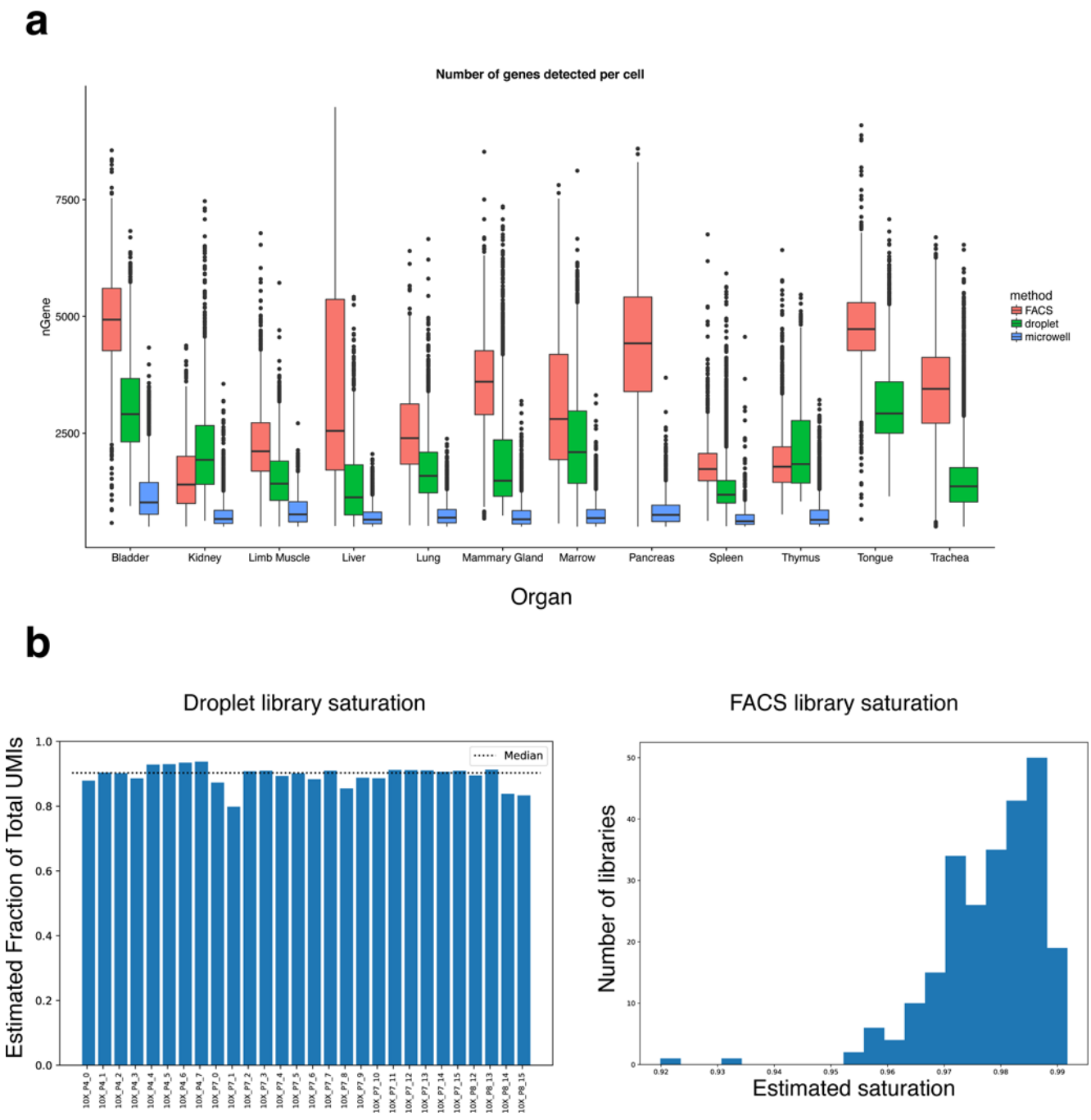




Extended Data Figure 4.

Graphical representation of cell ontology class representation.

a) FACS and b) microfluidic droplet datasets, colored by the relative amount of each cell type in each dataset.

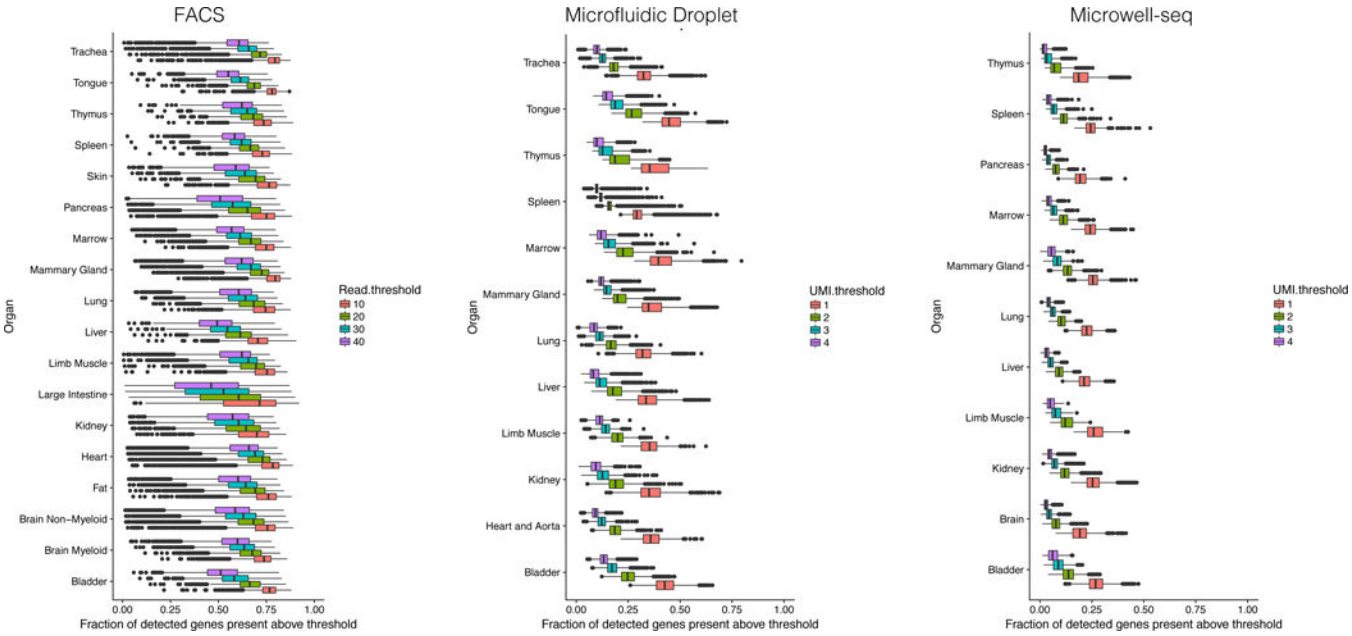


Extended Data Figure 5.

Methodological comparison of detected genes and library saturation.

a) The number of genes detected (threshold of > 0 reads or UMIs per cell) by FACS (red; n= 21,105 individual cells), microfluidic droplets (green; n = 55,032 individual cells), and microwell-Seq (blue; n= 25,891 individual cells) (Han *et al.*). b) Library saturation fraction for all microfluidic droplet libraries. Dotted horizontal line demarcates the median saturation (~ 0.9). c) Library saturation for all FACS libraries. Saturation was calculated using the

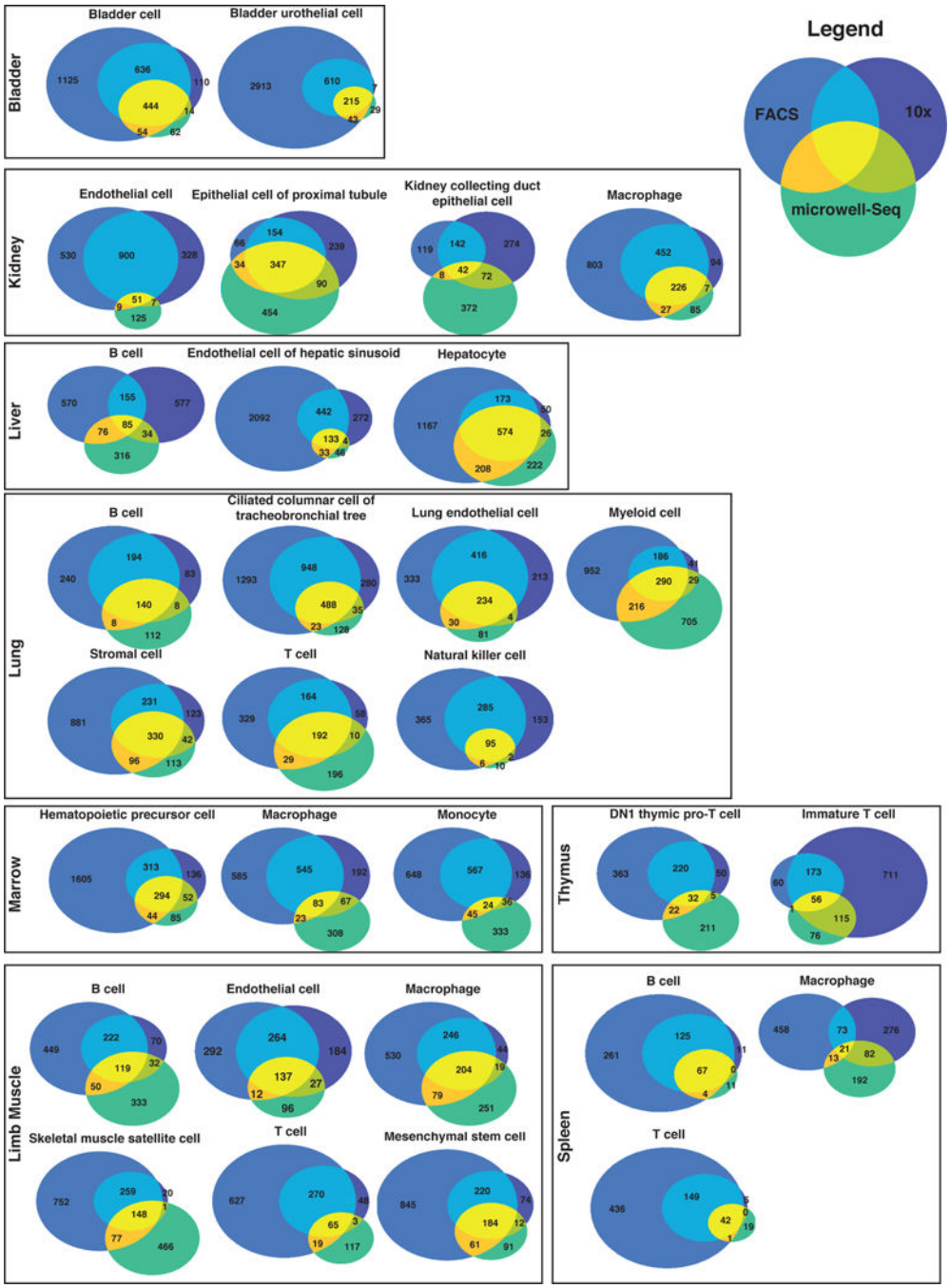
number of detected genes while downsampling the number of reads per library. Please refer to Supplementary Table 6 for summary statistics.



Extended Data Figure 6.

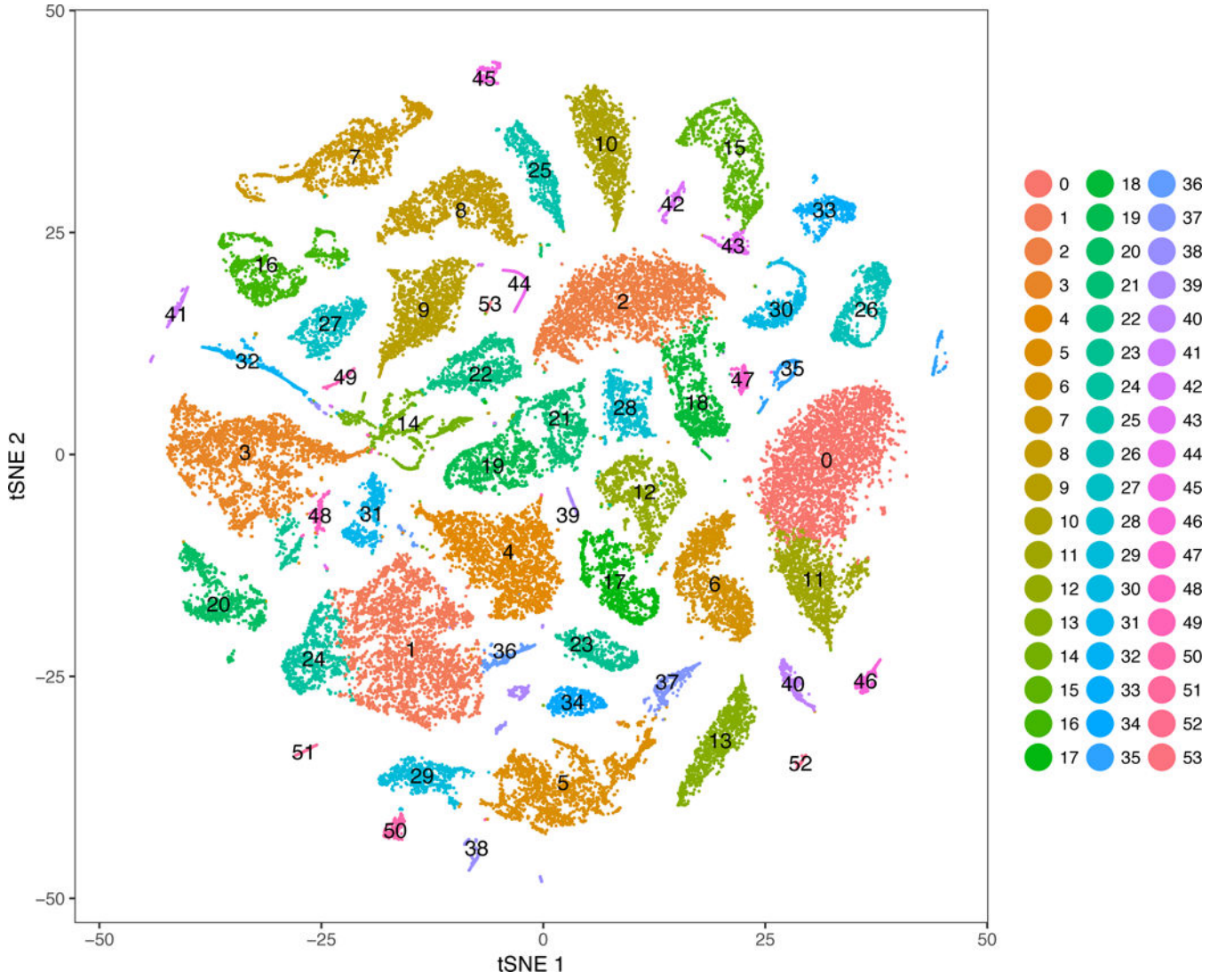
The number of detected genes decreases similarly across organs as the read or UMI threshold is increased.

Fraction of all detected genes (defined as > 0 reads or UMIs) for each cell, across all organs, detected at increasing read or UMI thresholds for FACS (left; n= 44,949 individual cells), microfluidic droplet (middle; n= 55,656 individual cells), and microwell-Seq (right; n= 28,372 individual cells). Please refer to Supplementary Table 6 for summary statistics.



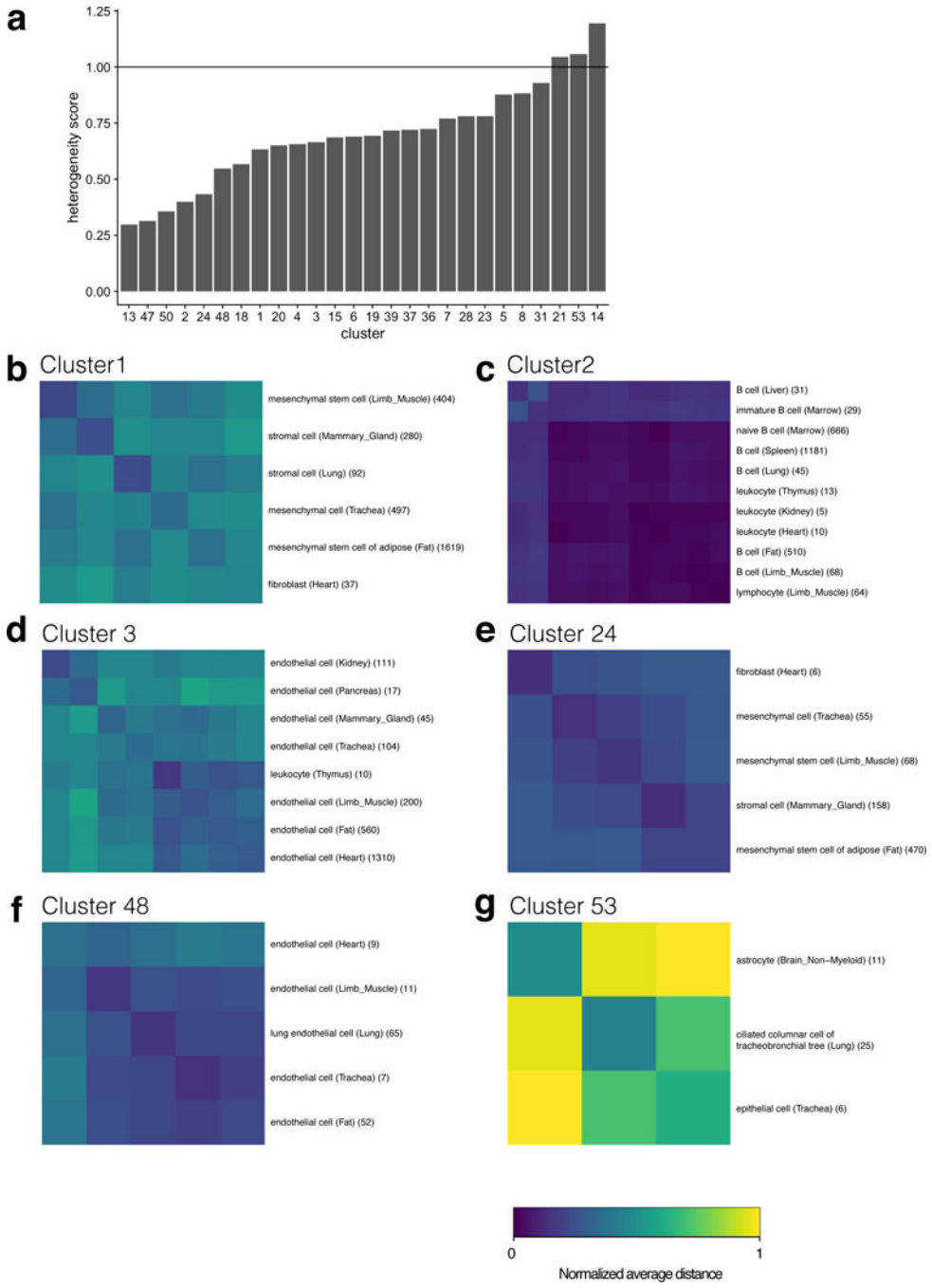
Extended Data Figure 7.

The number of differentially expressed genes for each cell type common between methods. Venn diagrams showing the overlap between differentially expressed genes for each common cell type across three methods (FACS, microfluidic droplet, microwell-Seq). Plotted data are provided in tabular form in Supplementary Table 2.



Extended Data Figure 8.

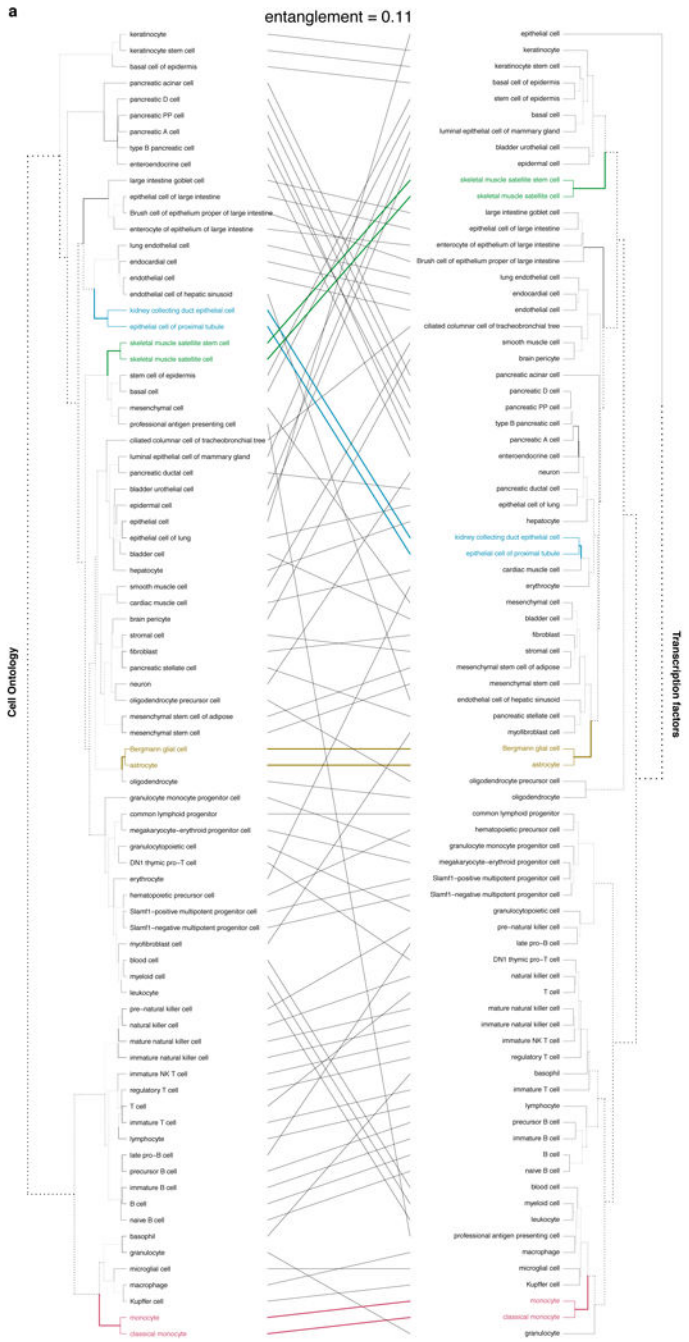
tSNE visualization of all FACS cells by cluster ID; n = 44,949 individual cells. Clusters are discussed in the text and further analyzed in Figure 3.

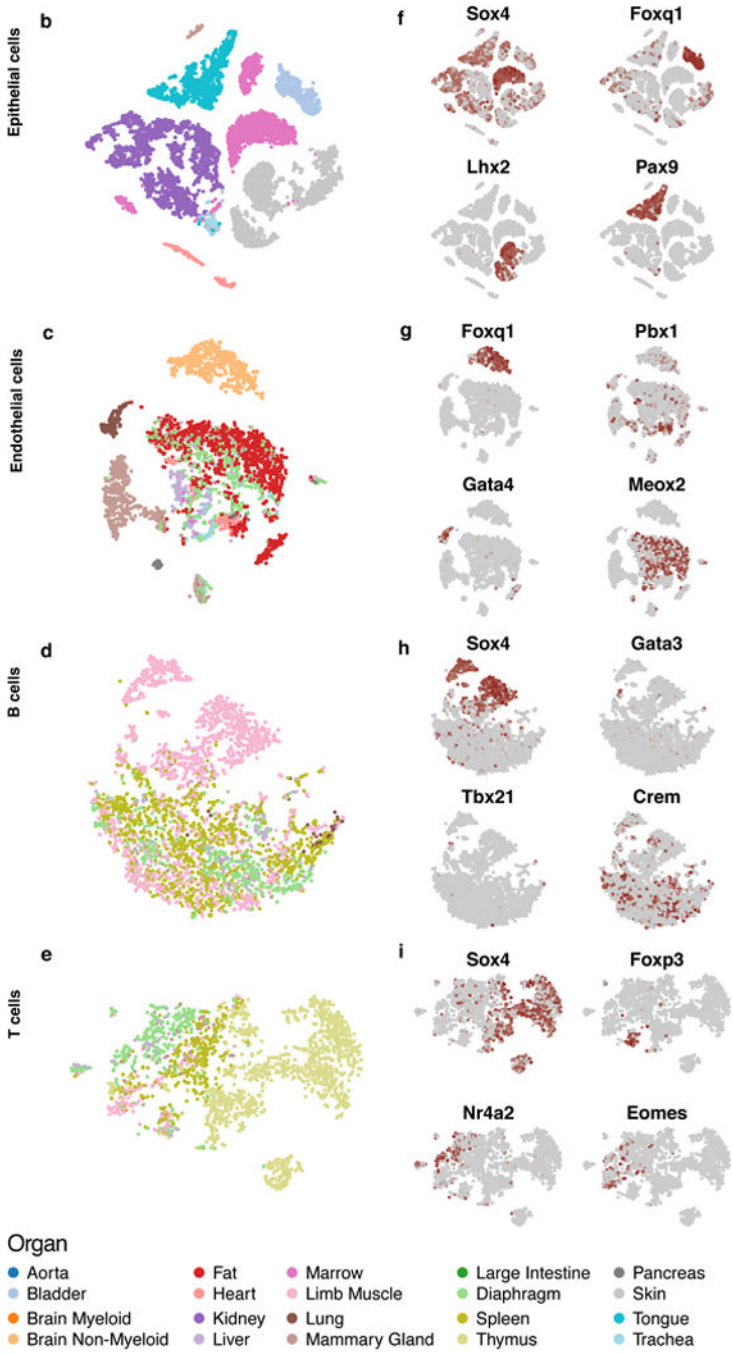


Extended Data Figure 9.

Metrics of cluster heterogeneity.

a) Barplot showing the heterogeneity score for each cluster containing multiple cell types. b-f) Heatmaps showing the average between-cell-type distances within select clusters, normalized so average distance between pairs of FACS cells is 1, clipped to a max of 1, for clusters a) 1, b) 2, c) 3, d) 24, e) 48, f) 53.



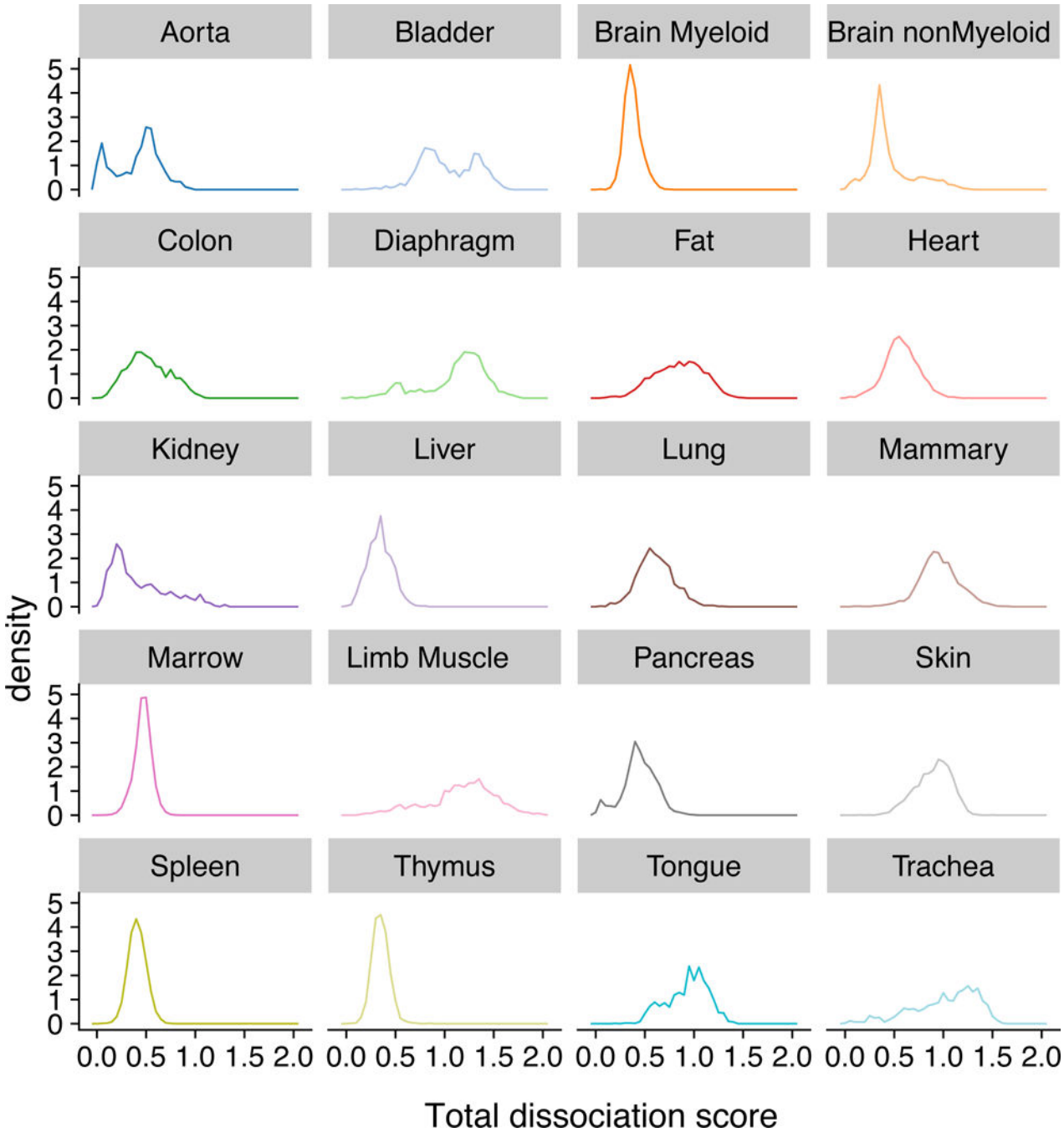


Extended Data Figure 10.

Transcription factor contribution to cell identity.

a) Tanglegram contrasting the dendrogram obtained using all expressed genes with one obtained using only the expression of TFs. The solid lines indicated segments that did not change position during the alignment between the two trees, while the dotted lines correspond to dendrogram branches re-ordered during the entanglement calculations. The colors indicate the branches for which the leaves are identical in both dendrograms. b-e) tSNE visualization of b) epithelial, c) endothelial, d) B-, e) T-cells colored by organ. f-i)

tSNE visualization of b) epithelial, c) endothelial, d) B-, e) T-cell expression of select TFs (grey/low to red/high). In b-i) n = 60 randomly selected cells for each cell type



Extended Data Figure 11.

Dissociation-induced gene expression scores for each organ analyzed with FACS. The dissociation score for each organ represents the magnitude of the first principal component of the 140 dissociation-associated genes from Van Der Brink *et al.* The y-axis shows probability density of the normalized histogram.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Nicholas Schaum[1], Jim Karkanias[2], Norma F. Neff[2], Andrew P. May[2], Stephen R. Quake[2,3,*], Tony Wyss-Coray[4,5,6,*], Spyros Darmanis[2,*], Joshua Batson[2], Olga Botvinnik[2], Michelle B. Chen[3], Steven Chen[2], Foad Green[2], Robert Jones[3], Ashley Maynard[2], Lolita Penland[2], Angela Oliveira Pisco[2], Rene V. Sit[2], Geoffrey M. Stanley[3], James T. Webber[2], Fabio Zanini[3], Ankit S Baghel[1], Isaac Bakerman[1,7,8], Ishita Bansal[2], Daniela Berdnik[4], Biter Bilen[4], Douglas Brownfield[9], Corey Cain[10], Michelle B. Chen[3], Steven Chen[2], Min Cho[2], Giana Cirolia[2], Stephanie D. Conley[1], Spyros Darmanis[2], Aaron Demers[2], Kubilay Demir[1,11], Antoine de Morree[4], Tessa Divita[2], Haley du Bois[4], Laughing Bear Torrez Dulgeroff[1], Hamid Ebadi[2], F. Hernán Espinoza[9], Matt Fish[1,11,12], Qiang Gan[4], Benson M. George[1], Astrid Gillich[9], Foad Green[2], Geraldine Genetiano[2], Xueying Gu[12], Gunsagar S. Gulati[1], Yan Hang[12], Shayan Hosseinzadeh[2], Albin Huang[4], Tal Iram[4], Taichi Isobe[1], Feather Ives[2], Robert Jones[3], Kevin S. Kao[1], Guruswamy Karnam[13], Aaron M. Kershner[1], Bernhard M. Kiss[1,14], William Kong[1], Maya E. Kumar[15,16], Jonathan Lam[12], Davis P. Lee[6], Song E. Lee[4], Guang Li[17], Qingyun Li[18], Ling Liu[4], Annie Lo[2], Wan-Jin Lu[1,9], Anoop Manjunath[1], Andrew P. May[2], Kaia L. May[2], Oliver L. May[2], Ashley Maynard[2], Marina McKay[2], Ross J. Metzger[19,20], Marco Mignardi[3], Dullei Min[21], Ahmad N. Nabhan[9], Norma F. Neff[2], Katharine M. Ng[3], Joseph Noh[1], Rasika Patkar[13], Weng Chuan Peng[12], Lolita Penland[2], Robert Puccinelli[2], Eric J. Rulifson[12], Nicholas Schaum[1], Shaheen S. Sikandar[1], Rahul Sinha[1,22,23,24], Rene V. Sit[2], Krzysztof Szade[1,25], Weilun Tan[2], Cristina Tato[2], Krissie Tellez[12], Kyle J. Travaglini[9], Carolina Tropini[26], Lucas Waldburger[2], Linda J. van Weele[1], Michael N. Wosczyna[4], Jinyi Xiang[1], Soso Xue[3], Justin Youngyunpipatkul[2], Fabio Zanini[3], Macy E. Zardeneta[6], Fan Zhang[19,20], Lu Zhou[18], Ishita Bansal[2], Steven Chen[2], Min Cho[2], Giana Cirolia[2], Spyros Darmanis[2], Aaron Demers[2], Tessa Divita[2], Hamid Ebadi[2], Geraldine Genetiano[2], Foad Green[2], Shayan Hosseinzadeh[2], Feather Ives[2], Annie Lo[2], Andrew P. May[2], Ashley Maynard[2], Marina McKay[2], Norma F. Neff[2], Lolita Penland[2], Rene V. Sit[2], Weilun Tan[2], Lucas Waldburger[2], Justin Y oungyunpipatkul[2], Joshua Batson[2], Olga Botvinnik[2], Paola Castro[2], Derek Croote[3], Spyros Darmanis[2], Joseph L. DeRisi[2,27], Jim Karkanias[2], Angela Oliveira Pisco[2], Geoffrey M. Stanley[3], James T. Webber[2], Fabio Zanini[3], Ankit S. Baghel[1], Isaac Bakerman[1,7,8], Joshua Batson[2], Biter Bilen[4], Olga Botvinnik[2], Douglas Brownfield[9], Michelle B. Chen[3], Spyros Darmanis[2], Kubilay Demir[1,11], Antoine de Morree[4], Hamid Ebadi[2], F. Hernán Espinoza[9], Matt Fish[1,11,12], Qiang Gan[4], Benson M. George[1], Astrid Gillich[9], Xueying Gu[12], Gunsagar S. Gulati[1], Yan Hang[12], Albin Huang[4], Tal Iram[4], Taichi Isobe[1], Guruswamy Karnam[13], Aaron M. Kershner[1], Bernhard M. Kiss[1,14], William Kong[1], Christin S. Kuo[9,11,21], Jonathan Lam[12], Benoit Lehallier[4], Guang Li[17], Qingyun Li[18], Ling Liu[4], Wan-Jin Lu[1,9], Dullei Min[21], Ahmad N. Nabhan[9], Katharine M. Ng[3], Patricia K. Nguyen[1,7,8,17], Rasika Patkar[13], Weng Chuan Peng[12], Lolita Penland[2], Eric J. Rulifson[12], Nicholas Schaum[1], Shaheen S. Sikandar[1], Rahul Sinha[1,22,23,24], Krzysztof Szade[1,25], Serena Y. Tan[22], Krissie Tellez[12], Kyle J. Travaglini[9], Carolina Tropini[26], Linda J. van Weele[1], Bruce M. Wang[13], Michael N. Wosczyna[4], Jinyi Xiang[1], Hanadie Yousef[4], Lu Zhou[18], Joshua

Batson[2], Olga Botvinnik[2], Steven Chen[2], Spyros Darmanis[2], Foad Green[2], Andrew P. May[2], Ashley Maynard[2], Angela Oliveira Pisco[2], Stephen R. Quake[2,3], Nicholas Schaum[1], Geoffrey M. Stanley[3], James T. Webber[2], Tony Wyss-Coray[4,5,6], Fabio Zanini[3], Philip A. Beachy[1,9,11,12], Charles K. F. Chan[28], Antoine de Morree[4], Benson M. George[1], Gunsagar S. Gulati[1], Yan Hang[12], Kerwyn Casey Huang[2,3,26], Tal Iram[4], Taichi Isobe[1], Aaron M. Kershner[1], Bernhard M. Kiss[1,14], William Kong[1], Guang Li[17], Qingyun Li[18], Ling Liu[4], Wan-Jin Lu[1,9], Ahmad N. Nabhan[9], Katharine M. Ng[3], Patricia K. Nguyen[1,7,8,17], Weng Chuan Peng[12], Eric J. Rulifson[12], Nicholas Schaum[1], Shaheen S. Sikandar[1], Rahul Sinha[1,22,23,24], Krzysztof Szade[1,25], Kyle J. Travaglini[9], Carolina Tropini[26], Bruce M. Wang[13], Kenneth Weinberg[21], Michael N. Wosczyna[4], Sean M. Wu[17], Hanadie Yousef[4], Ben A. Barres[18], Philip A. Beachy[1,9,11,12], Charles K. F. Chan[28], Michael F. Clarke[1], Spyros Darmanis[2], Kerwyn Casey Huang[2,3,26], Jim Karkanias[2], Seung K. Kim[12,29], Mark A. Krasnow[9,11], Maya E. Kumar[15,16], Christin S. Kuo[9,11,21], Andrew P. May[2], Ross J. Metzger[19,20], Norma F. Neff[2], Roel Nusse[9,11,12], Patricia K. Nguyen[1,7,8,17], Thomas A. Rando[4,5,6], Justin Sonnenburg[26], Bruce M. Wang[13], Kenneth Weinberg[21], Irving L. Weissman[1,22,23,24], Sean M. Wu[1,7,17], Stephen R. Quake[2,3], and Tony Wyss-Coray[4,5,6]

Affiliations

[1]Institute for Stem Cell Biology and Regenerative Medicine, Stanford University School of Medicine, Stanford, California, USA [2]Chan Zuckerberg Biohub, San Francisco, California, USA [3]Department of Bioengineering, Stanford University, Stanford, California, USA [4]Department of Neurology and Neurological Sciences, Stanford University School of Medicine, Stanford, California, USA [5]Paul F. Glenn Center for the Biology of Aging, Stanford University School of Medicine, Stanford, California, USA [6]Center for Tissue Regeneration, Repair, and Restoration, V.A. Palo Alto Healthcare System, Palo Alto, California, USA [7]Stanford Cardiovascular Institute, Stanford University School of Medicine, Stanford, California, USA [8]Department of Medicine, Division of Cardiology, Stanford University School of Medicine, Stanford, California, USA [9]Department of Biochemistry, Stanford University School of Medicine, Stanford, California, USA [10]Flow Cytometry Core, V.A. Palo Alto Healthcare System, Palo Alto, California, USA [11]Howard Hughes Medical Institute, USA [12]Department of Developmental Biology, Stanford University School of Medicine, Stanford, California, USA [13]Department of Medicine and Liver Center, University of California San Francisco, San Francisco, California, USA [14]Department of Urology, Stanford University School of Medicine, Stanford, California, USA [15]Sean N. Parker Center for Asthma and Allergy Research, Stanford University School of Medicine, Stanford, California, USA [16]Department of Medicine, Division of Pulmonary and Critical Care, Stanford University School of Medicine, Stanford, California, USA [17]Department of Medicine, Division of Cardiovascular Medicine, Stanford University, Stanford, California, USA [18]Department of Neurobiology, Stanford University School of Medicine, Stanford, CA USA [19]Vera Moulton Wall Center for Pulmonary and Vascular Disease, Stanford University School of Medicine, Stanford, California, USA [20]Department of Pediatrics, Division

of Cardiology, Stanford University School of Medicine, Stanford, California, USA [21]Department of Pediatrics, Stanford University school of Medicine, Stanford, California, USA [22]Department of Pathology, Stanford University School of Medicine, Stanford, California, USA [23]Ludwig Center for Cancer Stem Cell Research and Medicine, Stanford University School of Medicine, Stanford, California, USA [24]Stanford Cancer Institute, Stanford University School of Medicine, Stanford, California, USA [25]Department of Medical Biotechnology, Faculty of Biophysics, Biochemistry and Biotechnology, Jagiellonian University, Poland [26]Department of Microbiology & Immunology, Stanford University School of Medicine, Stanford, California, USA [27]Department of Biochemistry and Biophysics, University of California San Francisco, San Francisco, California USA [28]Department of Surgery, Division of Plastic and Reconstructive Surgery, Stanford University, Stanford, California USA [29]Department of Medicine and Stanford Diabetes Research Center, Stanford University, Stanford, California USA

Acknowledgements

We thank Sony Biotechnology for making an SH800S instrument available for this project. Some cell sorting/flow cytometry analysis for this project was done on a Sony SH800S instrument in the Stanford Shared FACS Facility. Some fluorescence activated cell sorting (FACS) was done with instruments in the VA Flow Cytometry Core, which is supported by the US Department of Veterans Affairs (VA), Palo Alto Veterans Institute for Research (PAVIR), and the National Institutes of Health (NIH).

References

1. Alberts B et al. Essential Cell Biology. (Garland Pub, 2014).
2. Guo G et al. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. Dev. Cell 18, 675–685 (2010). [PubMed: 20412781]
3. Dalerba P et al. Single-cell dissection of transcriptional heterogeneity in human colon tumors. Nat. Biotechnol. 29, 1120–1127 (2011). [PubMed: 22081019]
4. Thorsen T, Roberts RW, Arnold FH & Quake SR Dynamic pattern formation in a vesicle-generating microfluidic device. Phys. Rev. Lett. 86, 4163–4166 (2001). [PubMed: 11328121]
5. Macosko EZ et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161, 1202–1214 (2015). [PubMed: 26000488]
6. Klein AM et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161, 1187–1201 (2015). [PubMed: 26000487]
7. Ramsköld D et al. Full-length mRNA-seq from single-cell levels of RNA and individual circulating tumor cells. Nat. Biotechnol. 30, 777–782 (2012). [PubMed: 22820318]
8. Wu AR et al. Quantitative assessment of single-cell RNA-sequencing methods. Nat. Methods 11, 41–46 (2014). [PubMed: 24141493]
9. Treutlein B et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. Nature 509, 371–375 (2014). [PubMed: 24739965]
10. Enge M et al. Single-cell analysis of human pancreas reveals transcriptional signatures of aging and somatic mutation patterns. Cell 171, 321–330.e14 (2017). [PubMed: 28965763]
11. Halpern KB et al. Single-cell spatial reconstruction reveals global division of labour in the mammalian liver. Nature 542, 352–356 (2017). [PubMed: 28166538]
12. Haber AL et al. A single-cell survey of the small intestinal epithelium. Nature 551, 333–339 (2017). [PubMed: 29144463]
13. Villani A-C et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science 356, eaah4573 (2017).

14. Darmanis S et al. A survey of human brain transcriptome diversity at the single cell level. Proc. Natl. Acad. Sci. U.S.A. 112, 7285–7290 (2015). [PubMed: 26060301]
15. Gokce O et al. Cellular taxonomy of the mouse striatum as revealed by single-cell RNA-seq. Cell Rep 16, 1126–1137 (2016). [PubMed: 27425622]
16. Usoskin D et al. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. Nat. Neurosci. 18, 145–153 (2015). [PubMed: 25420068]
17. Zeisel A et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 347, 1138–1142 (2015). [PubMed: 25700174]
18. Li H et al. Classifying drosophila olfactory projection neuron subtypes by single-cell RNA sequencing. Cell 171, 1206–1220.e22 (2017). [PubMed: 29149607]
19. Bakken T et al. Cell type discovery and representation in the era of high-content single-cell phenotyping. BMC Bioinformatics 18, 559 (2017). [PubMed: 29322913]
20. Han X et al. Mapping the Mouse Cell Atlas by Microwell-Seq. Cell 172, 1091–1107.e17 (2018). [PubMed: 29474909]
21. Freytag S, Lonnstedt I, Ng M & Bahlo M Cluster headache: comparing clustering tools for 10X single cell sequencing data. bioRxiv (2017).
22. Holling TM, Schooten E & van Den Elsen PJ Function and regulation of MHC class II molecules in T-lymphocytes: of mice and men. Hum. Immunol. 65, 282–290 (2004). [PubMed: 15120183]
23. Reichardt J & Bornholdt S Statistical mechanics of community detection. Phys Rev E Stat Nonlin Soft Matter Phys 74, 016110 (2006).
24. van den Brink SC et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. Nat. Methods 14, 935–936 (2017). [PubMed: 28960196]
25. Alemany A, Florescu M, Baron CS, Peterson-Maduro J & van Oudenaarden A Whole-organism clone tracing using single-cell sequencing. Nature 556, 108–112 (2018). [PubMed: 29590089]
26. Cao J et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. Science 357, 661–667 (2017). [PubMed: 28818938]
27. Liu Z et al. Single-cell transcriptomics reconstructs fate conversion from fibroblast to cardiomyocyte. Nature 551, 100–104 (2017). [PubMed: 29072293]
28. Darmanis S et al. Single-cell RNA-seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. Cell Reports 21, 1399–1410 (2017). [PubMed: 29091775]
29. Kassambara A Practical guide to cluster analysis in R: unsupervised machine learning. CreateSpace Independent Publishing Platform: 2017.
30. Diaz-Uriarte R, & Alvarez de Andres S Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7, (2006).

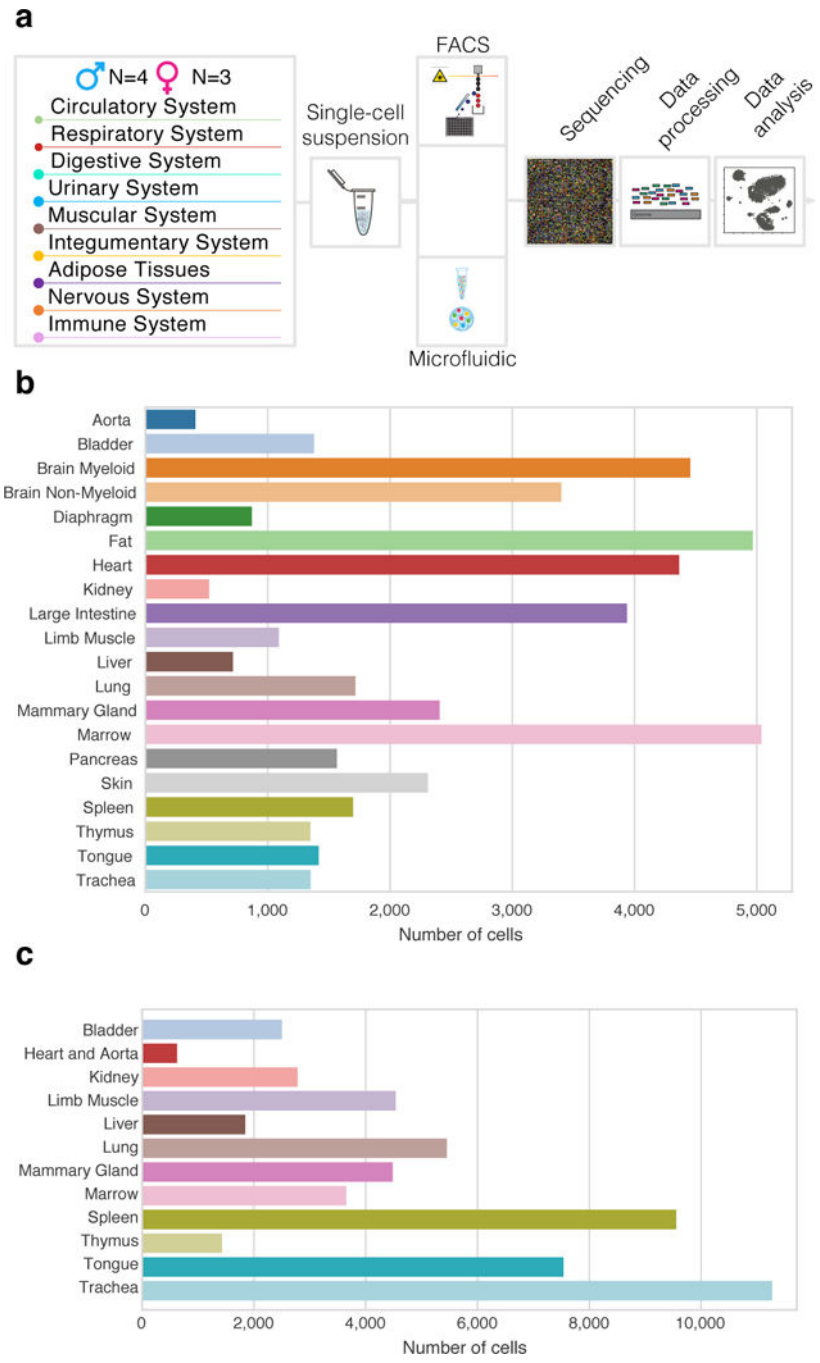


Figure 1.

Overview of Tabula Muris.

a) 20 organs from 4 male and 3 female mice were analyzed. After dissociation, cells were sorted by FACS and captured in microfluidic oil droplets for some organs. Cells were lysed, transcriptomes amplified and sequenced, reads mapped, and data analyzed. b) Barplot showing the number of sequenced cells prepared by FACS from each organ (n = 20 organ types). c) Barplot showing the number of sequenced cells prepared by microfluidic droplets from each organ (n = 12 organ types).

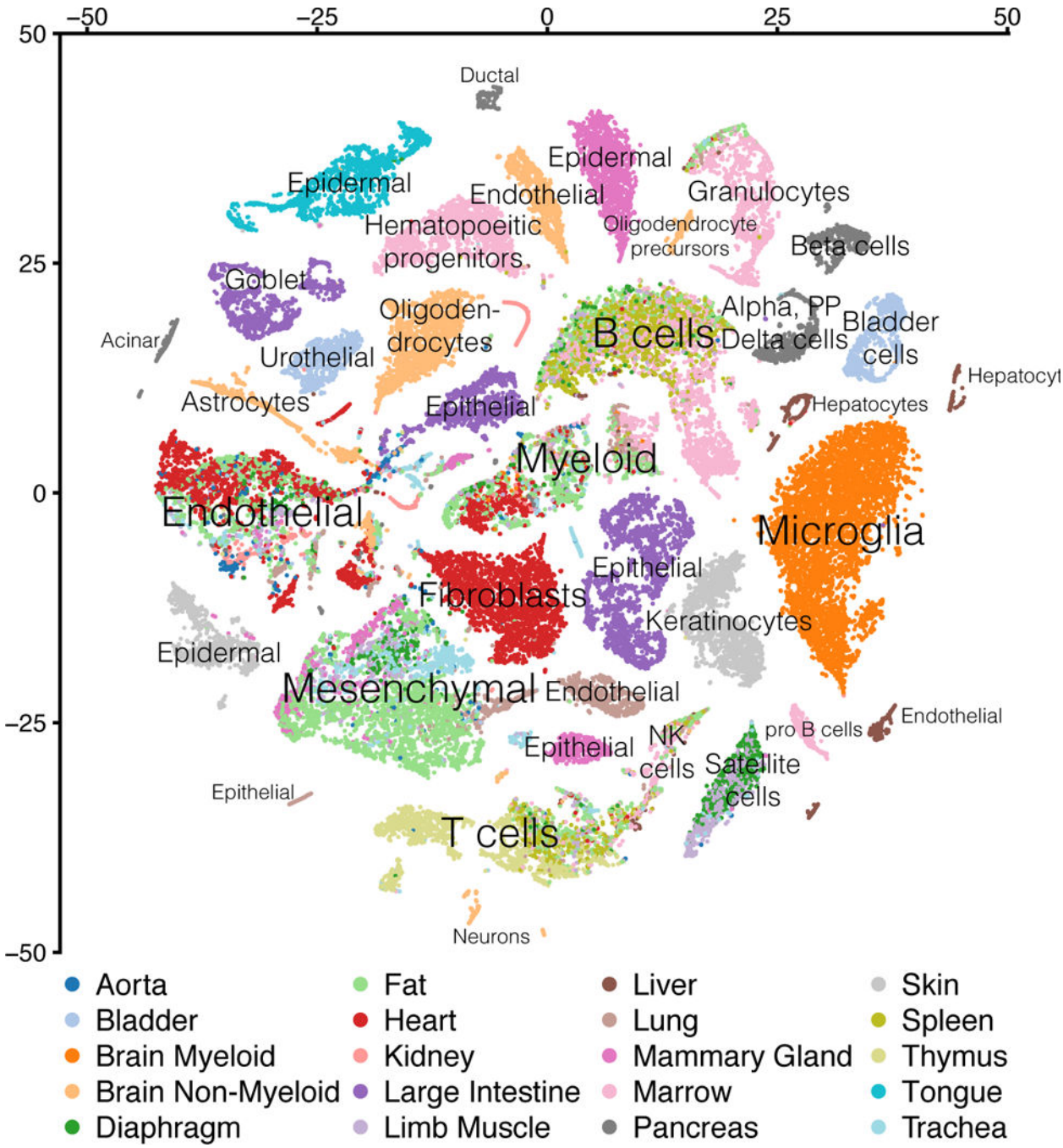


Figure 2. tSNE visualization of all FACS cells.

tSNE plot of all cells collected by FACS, colored by organ, overlaid with the predominant cell type composing each cluster; n = 44,949 individual cells.

Figure 3. Comparison of cell type determination.

Comparison of cell type determination as done by unbiased whole transcriptome comparison versus manual annotation of clusters by organ-specific experts. The x-axis represents clusters from Figure 2 with multiple organs contributing, while the y-axis represents manual expert annotation of clusters in an organ-specific fashion (Extended Data Fig. 1). The unbiased method discovers relationships between similar cell types found in different organs (highlighted regions); in particular it groups T cells from different organs into a single

cluster, B cells from different organs into a different single cluster, and endothelial cells from different organs into a single cluster.

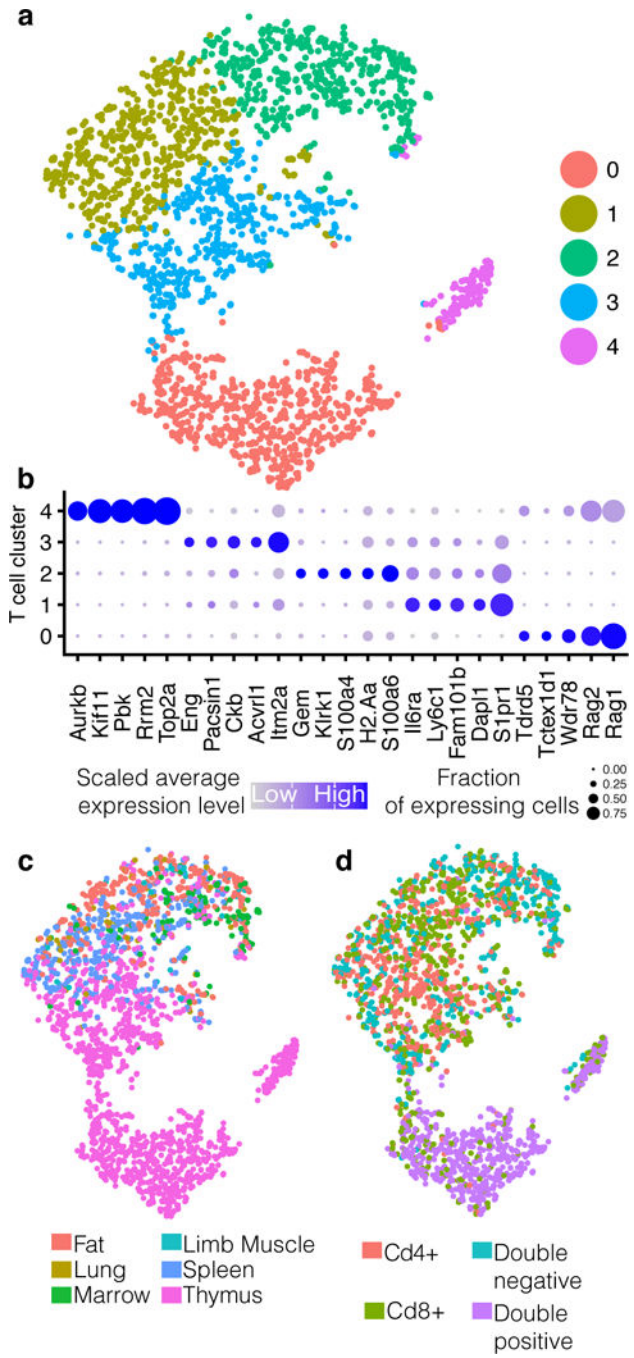


Figure 4. Analysis of all sorted T-cells.

a) tSNE plot of all T cells colored by cluster membership; n = 2847 individual cells. Five clusters were identified. b) Dotplot showing level of expression (color scale) and number of expressing cells (point diameter) within each cluster of T cells. c) tSNE plot of all T cells colored by organ of origin (Fat, Lung, Marrow, Limb Muscle, Spleen or Thymus); n = 2847 individual cells. d) tSNE plot of all T cells colored by classification of T cells to 4 categories based on expression of Cd4 and Cd8 ($Cd4^+$/ $Cd8^+$/ $Cd4^+Cd8^+$ / $Cd4^-Cd8^-$); n = 2847 individual cells.

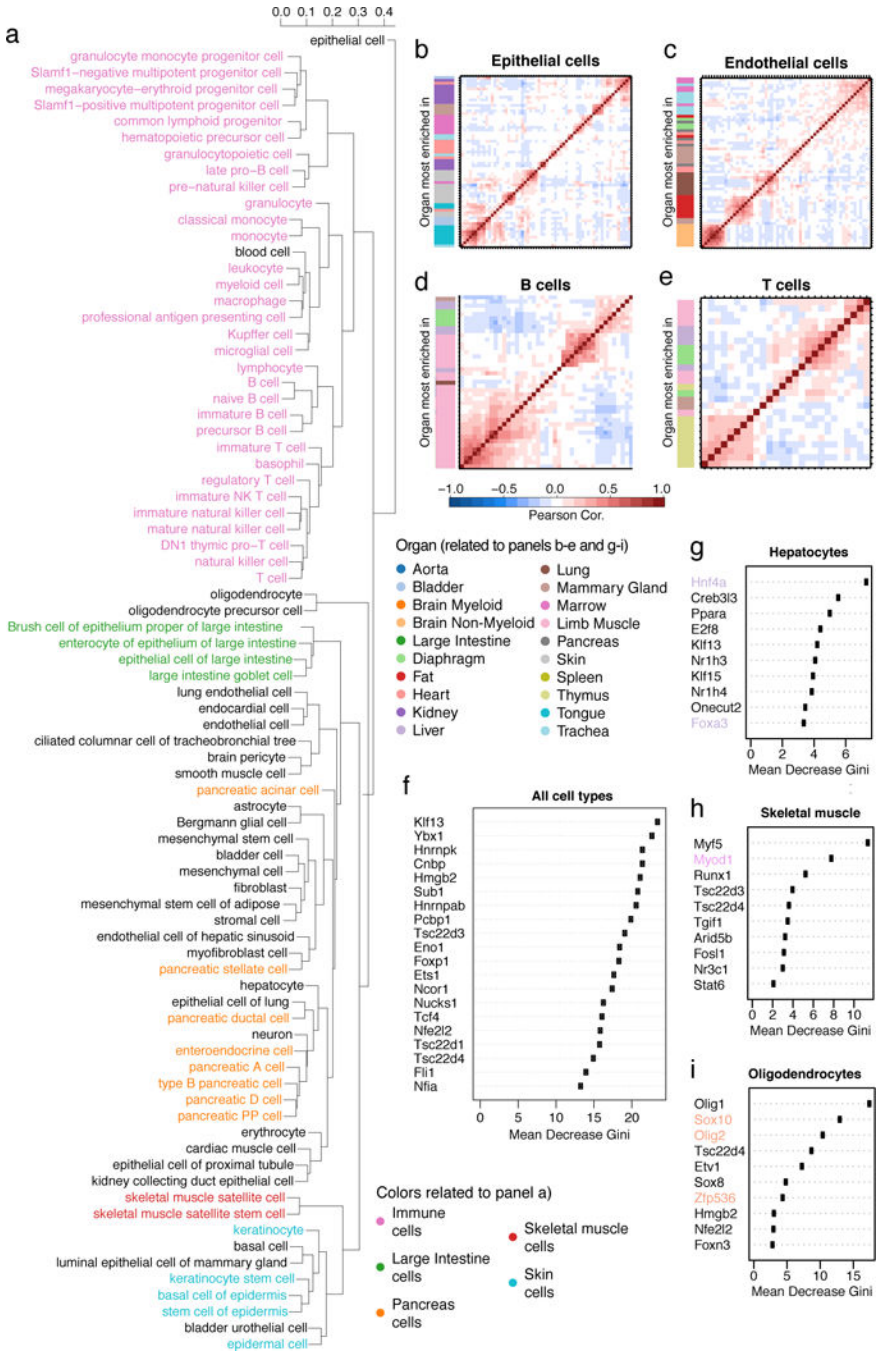


Figure 5. Transcription factor (TF) analysis.

a) Dendrogram of cell types constructed with only TFs. b-e) Correlograms of top organ-specific TFs for epithelial (b), endothelial (c), B- (d) and T- (e) cells. Row colors correspond to the organ of the most-enriched cell type; n=60 randomly selected cells for each cell type f) Top 20 TFs (mean Gini importance) of the random forest model when classifying all cell types. g-i) Top 10 TFs (mean Gini importance) of the random forest model when classifying each