Single-Cell Transcriptomic Analysis of mIHC Images via Antigen Mapping

Kiya W. Govek*, Emma C. Troisi*, Steven Woodhouse, and Pablo G. Camara#

Department of Genetics and Institute for Biomedical Informatics,
Perelman School of Medicine, University of Pennsylvania,
3700 Hamilton Walk, Philadelphia, PA 19104.

* These authors contributed equally to this work.

Correspondence to: pcamara@pennmedicine.upenn.edu

Histology provides a unique window into the cellular and molecular architecture of tissues and is a critical component of biomedical research and clinical practice. Highly-multiplexed immunohistochemistry[1-6] (mIHC) enables the routine staining and quantification of dozens of antigens in the same tissue section with single-cell resolution. However, the amount of cell types and states that can be simultaneously identified by mIHC is limited. In contrast, cells are finely disaggregated into distinct types in single-cell transcriptomic analyses but spatial information is lost. To bridge this gap, we developed an approach for enriching mIHC histology slides with single-cell RNA-seq data, building upon recent experimental procedures for augmenting single-cell transcriptomes with concurrent antigen measurements[7, 8]. Our approach, Spatially-resolved Transcriptomics via Epitope Anchoring (STvEA), increases the level of detail in histological analyses by enabling detection of subtle cell populations, spatial patterns of transcription, and cell-to-cell interactions. It provides an improvement in throughput, resolution, and simplicity with respect to existing spatially-resolved methods for simultaneous proteomics and transcriptomics. We demonstrate the utility of STvEA by uncovering the architecture of poorly characterized cell populations in the murine spleen using published mIHC images.

The most recent technologies for highly-parallelizable single-cell RNA-seq allow augmenting single-cell transcriptomes with concurrent protein measurements[7, 8]. For example, CITE-seq utilizes oligonucleotide-conjugated antibodies to combine multiplexed protein marker detection with unbiased transcriptome profiling of single cells[7]. We used the murine spleen as a test system to assess the feasibility of mapping CITE-seq data to mIHC images, since well-established antibody panels and high-quality mIHC data are readily available for this organ. One such high-resolution mIHC dataset has recently been generated using the CODEX technology[2].

CODEX employs an in-situ polymerization indexing procedure to measure the spatial distribution of a panel of protein markers with sub-micrometer resolution. We generated a high-quality CITE-seq dataset of the murine spleen using the same 30-antibody panel (Supplementary Table 1) and mice of the same age, sex, and strain as this CODEX dataset. In total, we profiled the transcriptome and antigen levels of 7,097 cells using CITE-seq, with a median of 819 expressed genes and 3,235 antibody-derived tag (ADT) unique molecular identifiers (UMIs) per cell. The median Spearman correlation among the observed expression of mRNAs and the proteins they code for was 0.32, consistent with previous CITE-seq studies[7]. We used single-cell variational inference (scVI)[9] to obtain a latent space representation of the mRNA data and clustered the cells in this space using an in-house consensus algorithm (see Online Methods). Our analysis found 17 clusters and no noticeable batch effects (Figs. 1a, Supplementary Fig. 1). We performed differential expression analysis to annotate the clusters based on the expression of known marker genes (Fig. 1b, Supplementary Table 2). Additionally, we utilized a spectral graph method[10, 11] to characterize the transcriptional heterogeneity that originates from the continuous and dynamic maturation processes occurring in the spleen (Fig. 1c, Supplementary Table 3). The results of this analysis allowed us to annotate the mRNA dataset beyond discrete clusters. Overall, we identified 30 cell populations (Fig. 1a), including T cells (helper, cytotoxic, effector memory, and regulatory), B cells (follicular, marginal zone, B-1, T1, and Notch2^{high}), natural killer cells, dendritic cells (plasmacytoid, conventional CD4 and CD8, pre-DCs), red pulp macrophages (AP-1^{high} and AP-1^{low}), monocyte-derived macrophages, monocytes, basophils, neutrophils (IL1^{high} and IL1^{low}, meta-myelocytes), plasma cells, Langerhans cells, erythroblasts (immature and mature), and CD47^{-} erythrocytes. These results represent a substantial increase in resolution with respect to previous single-cell RNA-seq atlases of the murine spleen[12-14] and comprise most of the known splenic cell populations[15, 16].

We noticed that most of the cell populations identified in the transcriptomic analysis were also localized in the protein expression space spanned by the ADT data (Supplementary Fig. 2). This observation indicates that small differences in cellular epitope levels are often representative of distinct transcriptomic states, even if those differences do not lead to discrete clusters in the protein expression space. Consequently, we reasoned that mapping the CODEX protein expression space into the CITE-seq protein expression space with high-resolution would allow us to survey the CODEX images for the cell populations identified in the transcriptomic analysis. To lessen the technical differences and facilitate the integration of the two spaces, we devised a common approach to background removal and normalization for CODEX and CITE-seq protein expression measurements (Fig. 2a). In each dataset, we modeled the distribution of protein levels using a two-component mixture model consisting of background and signal components (see Online Methods). Our approach led to improved and more consistent protein expression levels across the two datasets (Fig. 2a). We then employed a similar strategy to that of Stuart *et al.*[17] to consolidate the signal component of the two datasets into a common protein expression space (see Online Methods). To that end, we identified a set of anchors (pairs of cells in the CITE-seq and CODEX datasets that will be mapped to each other) by computing mutual nearest neighbors[18] in a common ambient space obtained via canonical correlation analysis (Fig. 2b). Anchors were filtered and weighted according to their degree of consistency with the protein and mRNA expression spaces. By looking at the CODEX neighbors of each CITE-seq cell in the consolidated protein expression space (Fig. 2b), we were able to identify cells in the mIHC images with a similar antigenic profile to those in the CITE-seq dataset. Using this approach, we predicted the spatial location of the cell populations identified in the transcriptomic analysis (Fig. 2c). STvEA correctly recapitulated the known spatial distribution of splenic cell populations, including the location of plasmacytoid dendritic cells (pDCs) in T cell zones and the positioning of CD4 conventional dendritic cells (cDCs) along the bridging channels that connect T cell zones and the red pulp[19] (Fig. 2c). We also observed that much of the transcriptional heterogeneity within B-2 cells was associated with distinct locations within B cell zones (Fig. 2c). The cell population assignments inferred for each cell in the CODEX dataset were consistent across CITE-seq replicates (median Pearson's correlation between population assignments r = 0.998, p-value < 10^{-10}), with the largest uncertainties occurring between AP-1^{high} and AP-1^{low} red pulp macrophages and between erythrocytes and erythroblasts (Fig. 2d). Additionally, the relative spatial distributions inferred by STvEA were reproducible across multiple spleens profiled with CODEX (Supplementary Fig. 3).

The mapping of single-cell transcriptomic data onto mIHC images provided by STvEA allows us to investigate the predicted spatial patterning of any gene in the mRNA dataset (Fig. 2e). To validate some of the spatially-resolved gene expression profiles predicted by STvEA, we performed multiplexed RNA fluorescent in situ hybridization[20] (FISH) of several marker genes identified in the differential expression analysis (Fig. 2e, Supplementary Figs. 4 and 5). Specifically, we carried out hybridizations for *Bhlhe41*, a transcriptional repressor highly expressed by B-1 cells[21] as they mature and migrate from B cell zones into the red pulp[22]; and *Il1b*, expressed by several subpopulations of cDCs, monocytes, macrophages, and neutrophils in the red pulp and T cell zones, but not expressed in B cell zones. In both cases, FISH correctly recapitulated the expression patterns predicted by STvEA (Fig. 2e, Supplementary Figs. 4 and

5), confirming the utility of our computational approach to label mIHC images by gene expression levels.

Characterizing cell-to-cell interactions within the context of tissues is a key step towards understanding cell function. Having inferred the transcriptome and cell type of individual cells in mIHC images enabled us to survey candidate cell-to-cell interactions with cellular resolution. We devised a graph-based approach for assessing the spatial co-localization of cell populations identified in the transcriptomic analysis while accounting for mapping uncertainties (see Online Methods). Significant co-localization patterns recapitulated the well-established immune cellular architecture of the spleen, partitioned into red pulp, B cell zones, and T cell zones (Fig. 3a). T cells, pDCs, and CD4 cDCs were recurrently in close proximity within T cell zones. Similarly, red pulp macrophages, erythrocytes, neutrophils, and monocytes were recurrently in close proximity within the red pulp. In addition, several cell populations showed co-localization patterns that spanned multiple splenic compartments (Fig. 3a). Specifically, CD4 cDCs appeared recurrently in close proximity with T cells in T cell zones and with NK cells in the red pulp (Fig. 3a). These inferred relations were reproducible across multiple spleens profiled with CODEX (Fig. 3b, Pearson's correlation coefficient between significance levels, $r \geq 0.98$).

To identify molecular cues that potentially mediate the crosstalk between splenic cell populations, we compared differentially expressed genes to a database of receptor-ligand interactions[23]. This analysis identified 67 significant candidate interactions based on the expression of genes encoding for ligands and receptors by one or more cell populations (CellPhoneDB p-value ≤ 0.05, Supplementary Table 4). However, many of these interactions would not be functional if ligand- and receptor-expressing cells are not in close proximity. Thus, we subsequently restricted our analysis to pairs of ligand and receptor genes with significantly co-localized expression in the tissue section. Overall, we detected 29 significant receptor-ligand interactions based on the proximity of ligand- and receptor-expressing cells (Benjamini-Hochberg adjusted q-value ≤ 0.05, median distance 5 μm, Supplementary Table 4). The results of this analysis suggest a specialization of resident red pulp macrophages in regulating the homeostasis of humoral innate immune responses in the spleen (Fig. 3c). We identified the expression of several cues related to the positive regulation of these responses. Complement component 1q (C1q) can mediate phagocytosis of apoptotic cells by calreticulin/LRP1 receptor stimulation[24]. We observed the co-localized expression of genes encoding for C1q subunits and the receptor LRP1 in both red pulp and monocyte-derived macrophages (Fig 3c and Supplementary Table 4). However, the expression of C1q genes was 12-fold higher in red pulp macrophages than in monocyte-derived macrophages (p-value < 10^{-10}), suggesting the specialization of the former in modulating C1q-dependent phagocytosis. Red pulp macrophages also displayed high expression levels of the *Hebp1* gene, which encodes for the precursor of the F2L peptide that activates and chemo-attracts neutrophils by binding to the formyl peptide receptor Fpr2[25]. Consistent with a function of red pulp macrophages in orchestrating neutrophil activation, *Hebp1*-expressing red pulp macrophages were in close spatial proximity to *Fpr2*-expressing neutrophils (Fig. 3c, Supplementary Table 4). In contrast, monocyte-derived macrophages displayed low expression levels of *Hebp1* (fold-change with respect to red pulp macrophages = 0.1, p-value < 10^{-10}). These macrophages instead expressed annexin A1 (*Anxa1*) (Fig. 3c, Supplementary Table 4), an anti-inflammatory agonist of the neutrophilic receptor Fpr1[26]. Additionally, we observed high expression levels of insulin-like growth factor 1 (*Igf1*) in red pulp macrophages. Various works have put forward the role of human IGF-1 in stimulating the activation and chemokinesis of basophils through the IGF-1R receptor[27-30]. Consistent with this hypothesis, our analysis revealed red pulp macrophages were recurrently adjacent to *Igf1r*-expressing basophils (Fig. 3c, Supplementary Table 4). The expression of *Igf1* in monocyte-derived macrophages was substantially lower (fold-change with respect to red pulp macrophages = 0.1, p-value < 10^{-10}), indicating the specialization of resident red pulp macrophages in IGF-1 signaling. Hence, taken together these results suggest important differences in the immuno-regulatory function of red pulp and monocyte-derived macrophages in the murine spleen. More broadly, they show the utility of spatially-resolved expression data in the study of cell-to-cell signaling interactions.

Several platforms have been recently developed for highly-multiplexed spatially-resolved transcriptomics with single-cell resolution[31-38]. However, these experimental approaches are technically challenging and costly to implement, and do not permit simultaneous measurements of RNA and protein levels. Although a platform for highly-multiplexed spatial profiling of proteins and RNA has been recently proposed, profiling more than a small number (~10) of cells in a tissue section with this system is currently impractical and cost-prohibitive[39]. STvEA provides a simple computational approach for enriching mIHC images with single-cell transcriptomic data using widely-accessible commercial platforms. Its throughput and resolution allows for the characterization of the expression levels of thousands of genes for tens of thousands of cells in a tissue slide. Furthermore, it can leverage existing CITE-seq and mIHC datasets. We have implemented STvEA as open source software available to the entire community (see Online Methods). The results of our murine spleen study can be accessed interactively through an online database (see Online Methods). We expect these resources to be of great utility for future studies of the cellular and molecular architecture of healthy and diseased tissues.

Methods

Mouse handling

All animal work was approved by and carried out in compliance with the animal welfare regulations defined by the University of Pennsylvania International Animal Care and Use Committee (IACUC). 15 week old female BALB/cJ (Stock #000651) mice were acquired from The Jackson Laboratory (Bar Harbor, ME). Mice were allowed to age at the University of Pennsylvania Small Animal Facility until they reached approximately 9 months, at which point they were euthanized using CO2 followed by cervical dislocation.

Tissue dissection and preparation of splenic single-cell suspensions

Spleens were removed from mice and mechanically dissociated with a syringe plunger over a 40 um strainer while being washed with 5 ml of PBS + 10% fetal calf serum. Suspension were centrifuged briefly to pellet cells. Red blood cells were lysed with an RBC lysis buffer (155 mM NH4Cl, 12 mM NaHCO3, 0.1 mM EDTA) for 5 minutes and centrifuged again. 2 million cells from the resulting pellet were re-suspended in staining buffer (2% BSA, 0.01% Tween in PBS), and subsequently incubated with the antibody panel as described below (see “Cell Staining”).

CITE-seq antibody conjugation and panel preparation

Antibodies were conjugated to 5’ amino-modified, HPLC-purified CITE-seq oligonucleiotides purchased from Integrated DNA Technologies. Antibodies were concentrated to 1 mg/ml in PBS pH 7.4 using 50 kDa cutoff spin columns (UFC505024, Millipore). Oligonucleiotides were resuspended to 1 mg/ml in 1x PBS pH 7.4 and were subsequently cleaned as suggested in the CITE-seq protocol. In brief, oligos were heated at 85C and centrifuged at 17,000g to pellet any debris. For each antibody, 100 ug of antibody and 100 ug of oligo were conjugated using the Thunder-Link PLUS Oligo Conjugation System (SKU: 425-0300, Expedeon). All conjugates were cleaned as described in the CITE-seq protocol and resuspended to their final concentration in the Antibody Resuspension Buffer provided with the kit, with the exception of CD16/32, which was resuspended in 1x PBS. Successful conjugation was validated by running 1 ug of each conjugate on a 2% agarose gel which was subsequently stained with Sybr Gold (S11494, Thermo Fisher Scientific).

To prepare the panel, 1.5 ul of each antibody-oligo conjugate (except CD16/32) were combined in PBS and centrifuged in a 50 kDa cutoff column. After washing, the cleaned panel was recovered by flipping the column upside down and centrifuging. The cleaned panel was resuspended in staining buffer.

Single-cell CITE-seq library preparation and sequencing

1.5 ug of the CD16/32 antibody-oligo conjugate was incubated with the single cell suspension for 10 minutes in place of the mouse seroblocker suggested in the CITE-seq protocol. The remaining 29 antibodies were then added to the cell suspension and incubated on ice. After incubation, cells were washed thoroughly, counted on a hemocytometer, and loaded into the 10x Chromium platform (10x Genomics) for single-cell library preparation. Cells were loaded at 1,200 cells/ul. Only samples with >80% cell viability were used, profiling a total of 2 mouse spleens. cDNA libraries were prepared following the standard CITE-seq and 10x protocols. The resulting ADT and mRNA libraries were combined at a 1:9 ratio and sequenced with an Illumina HiSeq 2500 at the Center of Applied Genomics, Children's Hospital of Philadelphia.

Multiplexed RNA FISH of splenic tissue sections

Whole spleens were removed from euthanized mice and immediately submerged in 4% paraformaldehyde for 5.5 hours. They were then cryoprotected in a 30% sucrose/70% fixative solution at 4°C until the tissue sank (approximately overnight, ~16 hours). The tissue was embedded in OCT cryostat sectioning medium (OCT Compound, Sakura Finetek Inc, Supp. No. 4583) on dry ice and frozen at -80°C. Tissue was cut using a cryostat at -20°C into 10 um-thick sections and frozen again at -80°C. Tissue was used for microscopy within 6 months of fixation and cryoprotection.

RNA fluorescence in situ hybridization experiments were carried out with the RNAscope Multiplex Fluorescence Reagent Kit v2 (Advanced Cell Diagnostics, Hayward, CA, USA, Cat. No. 323100). The RNAscope Assay for fixed frozen samples was followed per the manufacturer's protocol with the following two modifications: the post-fix incubation was carried out with 4% PFA at room temperature for 90 minutes and manual target retrieval with a 5 minute sample incubation was performed instead of the steamer method. Probes for mouse *Bhlhe41* and *Il1b* (Advanced Cell Diagnostics, Cat. No. 467431 and 316891) were hybridized with Opal 520 (Akoya Biosciences, Cat. No. FP1487001KT), and probes for mouse *Cd79a* (Advanced Cell Diagnostics, 460181-C2) was hybridized with Opal 570 (Akoya Biosciences, Cat. No. FP1488001KT). Both dyes were diluted 1:1500 with TSA buffer provided by the RNAscope kit. Channel 2 was diluted in channel 1 1:50 as suggested in the RNAscope protocol. All incubations were carried out using a Stratagene PersonalHyb hybridization oven. Sequential sections were processed alongside the positive and negative controls provided by the RNAscope kit. Immunofluorescence images were acquired using a Leica TCS SP8 Multiphoton confocal microscope.

Single-cell CITE-seq processing

We used Cell Ranger to de-multiplex, map to the mouse reference genome (mm10), and count UMIs in the mRNA libraries, and CITE-seq-Count to count UMIs in the ADT libraries. We filtered out cells with more than 10% UMIs from mitochondrially-encoded genes or less than 1,200 mRNA UMIs in total. We used scVI to infer a lower dimensional latent space for visualization and clustering of the mRNA expression data. scVI uses a neural network to fit a zero-inflated negative binomial model to represent the technical variation in scRNA-seq data and create a latent space. We inferred an 18-dimensional latent space representation for the expression data of all genes expressed in at least 15 cells (training size = 0.75, number of epochs = 400, learning rate = 1×10^{-3}). The dimensionality of the latent space was empirically chosen based on the stability of the resulting representations and was consistent with the elbow of the scree plot. To visualize the mRNA expression, we further reduced the latent space to 2 dimensions using UMAP with Pearson's correlation distance.

Clustering and differential expression analysis of single-cell mRNA data

We clustered the cells in the latent space using HDBSCAN and an in-house consensus algorithm. Prior to clustering, we used UMAP to establish a metric in the 18-dimensional latent space, as suggested by the UMAP Python documentation. Then we scanned across the min_cluster_size and min_sample parameters of HDBSCAN (min_cluster_size ∈ $\{5,9,13,17\}$, min_sample ∈ $\{10,13,16,19,22,25,28,31,34,37\}$) and used cluster-based similarity partitioning to build a consensus matrix,

$$M_{ij} = \sum_{s \in S} s_{ij} \; , \qquad\qquad s_{ij} = \begin{cases} 0, & c_{si} = c_{sj} \\ 1, & \text{otherwise} \end{cases}$$

where S is the set of indicator functions for all parameter configurations that gave rise to clusters with a silhouette score > 0.114, and c_{si} is the cluster ID of cell i in s. We used this consensus matrix as a dissimilarity matrix among cells to produce a consensus clustering using average linkage agglomerative clustering (inconsistent value ≤ 0.1).

We ran edgeR's general linear model (GLM) on the mRNA count data to identify differentially expressed genes between each cluster and all the other cells (fold change threshold > 2).

Laplacian score analysis of single-cell mRNA data

We utilized the Laplacian score to more accurately annotate the mRNA data by identifying genes that have expression patterns within a cluster which cannot be explained by random variation. For each cluster, we built a graph where nodes represent cells and edges connect pairs of cells that are within ε distance, as defined by Pearson's correlation in the latent space. We took ε to be given by the median pairwise distance among cells. For large clusters, we randomly sampled 1,000 cells. The Laplacian score of a gene with expression vector f is defined as

$$L = \frac{\tilde{f}^T \cdot L \cdot \tilde{f}}{\tilde{f}^T \cdot D \cdot \tilde{f}}$$

where

$$\tilde{f} = f - \frac{f^T \cdot D \cdot \mathbf{1}}{\mathbf{1}^T \cdot D \cdot \mathbf{1}}\mathbf{1}, \qquad D = \mathrm{diag}(A \cdot \mathbf{1}), \qquad \mathbf{1} = [1, \dots, 1]^T, \qquad L = D - A$$

and A is the adjacency matrix of the graph. We computed the Laplacian score of the $\log(1 + TPM \cdot 10^{-2})$ expression values for all genes expressed in at least 2% and at most 90% of the cells. To assess the significance of the Laplacian score as compared to random variation, we performed a permutation test by randomizing the cell labels 1,000 times.

Normalization of ADT libraries

We fit the distribution of ADT counts for each antibody with a two-component negative binomial mixture model,

$$Prob(r = k_{hi}) \sim b_h \cdot NB\left(k_{hi}\,;\, r_h^{(1)}, p_h^{(1)}\right) + (1 - b_h) \cdot NB\left(k_{hi}\,;\, r_h^{(2)}, p_h^{(2)}\right)$$

where k_{hi} represents the observed number of ADT UMIs for antigen h in cell i, and the mixing parameter b_h represents the probability of a measurement of antigen h actually coming from the background. Upon fitting the model using least-squares estimation, we filtered out the background component of the data by considering the matrix

$$q_{hi}^{CITEseq} \equiv \frac{Prob(r \leq k_{hi} \mid k_{hi} \in signal)}{\sum_g k_{gi}}$$

where the signal component is defined as the component of the mixture model with higher median. We performed batch correction on the $q_{hi}^{CITEseq}$ values using the mutual nearest neighbors approach of Haghverdi *et al.*[18] and rescaled the resulting values to be in the [0,1] interval.

We did not consider CD169 in our analysis as it was showing expression in cell populations other than macrophages, possibly reflecting a lack of affinity of the conjugated antibody. In addition, the unimodal distribution of ERTR7 expression values was consistent with the fact that we did not capture stromal cells in our CITE-seq dataset (possibly because of the use of a non-enzymatic dissociation procedure). We therefore assigned all cells to the background component and set $q_{\mathrm{ERTR7},i}^{CITEseq} = 0$.

Processing of CODEX data

We considered the segmented and compensated CODEX data of the three wild-type mice profiled by Goltsev *et al.*[2]. We filtered out artifacts using a similar gating strategy to that of the CODEX protocol. We removed cells smaller than 1,000 or larger than 25,000 voxels. We then identified maximum and minimum cutoffs for blank channels by plotting the expression of one blank channel versus another, as described in the CODEX protocol. We removed cells with intensities above the upper cutoffs in any of the blank channels or below the lower cutoffs in all of the blank channels. Our cutoffs fell around the 99.5 and 0.2 percentiles respectively. However, we checked that small variations of the specific values did not greatly affect the number of cells removed.

Normalization of CODEX data

We normalized the processed CODEX data by the total levels in each cell,

$$\widehat{M}_{hi} = \frac{M_{hi}}{\sum_g M_{gi}}$$

where M_{hi} is the level of antigen h in cell i before normalization. After this process, antigen levels are well approximated by a two-component Gaussian mixture model,

$$Prob(r = \widehat{M}_{hi}) \sim a_h \cdot \mathcal{N}\left(\widehat{M}_{hi}\,;\, \mu_h^{(1)}, \sigma_h^{(1)}\right) + (1 - a_h) \cdot \mathcal{N}\left(\widehat{M}_{hi}\,;\, \mu_h^{(2)}, \sigma_h^{(2)}\right)$$

where the lower (higher) median Gaussian corresponds to the background (signal), and the mixing parameter a_h represents the probability of a measurement of antigen h actually coming from the background. Upon fitting the model to the data using the EM algorithm for maximum likelihood estimation, we filtered out the background component of the data by considering the probabilities

$$p_{hi} \equiv Prob(r \leq \widehat{M}_{hi} \mid \widehat{M}_{hi} \in signal)$$

in subsequent analysis.

Mapping of CODEX data into CITE-seq

We mapped the inferred CODEX probabilities p into the CITE-seq space $q^{CITEseq}$ using a modified version of the general strategy proposed by Stuart *et al.*[17]. Specifically, we identified a set of anchors using a mutual nearest neighbors approach with $k_{\text{anchor}} = 19$. We found the nearest neighbors using Euclidean distance in a common 29-dimensional space obtained by canonical correlation analysis (CCA). We then filtered out anchors that do not preserve the structure of the original protein space. For that purpose, we kept only those for which the CODEX cell in the anchor was within the $k_{\text{filter}} = 99$ nearest CODEX cells to the CITE-seq cell in the anchor, or vice versa, as measured by Pearson's correlation distance between p and $q^{CITE-seq}$.

Cells in the CODEX dataset were aligned into the CITE-seq protein space using the following transformation

$$q_{hi}^{CODEX} \equiv p_{hi} + \sum_{(j_1,j_2)\in\mathcal{A}_i} (q_{hj_1}^{CITEseq} - p_{hj_2}) \cdot w_{(j_1,j_2),i}$$

where $\mathcal{A}_i$ is the set of $k_{\text{weight}} = 99$ anchors (j_1, j_2) with smallest Pearson's correlation distance between cell j_2 and cell i, and $w_{(j_1,j_2),i}$ are weights specifying the effect size of anchor (j_1, j_2) on the CODEX cell i based on both mRNA and protein data,

$$w_{(j_1,j_2),i} = \frac{1 - e^{-d_{ij_2} s_{j_1 j_2}/c}}{\sum_{(j_1,j_2)\in\mathcal{A}_i} 1 - e^{-d_{ij_2} s_{j_1 j_2}/c}}$$

In this equation, d_{ij_2} denotes Pearson's correlation distance between the vectors $\vec{p}_i$ and $\vec{p}_{j_2}$ (with components p_{hi} and p_{hj}, respectively), and c is a parameter specifying the width of the Gaussian kernel. The number of shared neighbors between the two anchor cells, $s_{j_1 j_2}$, is defined as

$$s_{j_1 j_2} = \left|\mathcal{N}_{j_1}^{CITEseq} \cap \mathcal{N}_{j_2}^{CITEseq}\right| + \left|\mathcal{N}_{j_1}^{CODEX} \cap \mathcal{N}_{j_2}^{CODEX}\right|$$

where $\mathcal{N}_{j_1}^{CITEseq}$ is the set of nearest CITE-seq cells to cell j_1 in the mRNA latent space, $\mathcal{N}_{j_2}^{CITEseq}$ is the set of nearest CITE-seq cells to cell j_2 in the CCA space, $\mathcal{N}_{j_1}^{CODEX}$ is the set of nearest CODEX cells to cell j_1 in the CCA space, and $\mathcal{N}_{j_2}^{CODEX}$ is the set of nearest CODEX cells to cell j_2 in the CCA space. As before, distances in the mRNA and CCA spaces were measured using Pearson's correlation and Euclidean distance respectively. In all cases, the number of nearest neighbors was chosen to be $k_{\text{score}} = 79$. The values $s_{j_1 j_2}$ were scaled such that the 0.9 quantile is at 1 and the 0.01 quantile is at 0, and values above or below these quantiles were set to 1 or 0 respectively.

To be able to transfer quantities between the CITE-seq and CODEX datasets, we then built a $\mathcal{M}^{CITEseq\to CODEX}$ transfer matrix,

$$\mathcal{M}_{ij}^{CITEseq\to CODEX} \equiv \begin{cases} \dfrac{e^{-\tilde{d}_{ij}/c}}{\sum_{r:\, j\in\mathcal{N}_r^{CODEX}} e^{-\tilde{d}_{rj}/c}} & \text{iff} \quad j \in \mathcal{N}_i^{CODEX} \\ 0 & \text{iff} \quad j \notin \mathcal{N}_i^{CODEX} \end{cases}$$

where $\tilde{d}_{ij}$ denotes Pearson's correlation distance between the vectors $\vec{q}_i$ and $\vec{q}_j$ (with components q_{hi}^A and q_{hj}^B, respectively), and c is a parameter that specifies the width of the Gaussian kernel. The set $\mathcal{N}_i^{CODEX}$ contains the nearest CODEX cells to the CITE-seq cell i as measured by $\tilde{d}_{ij}$, where $k_{\text{transfer}} \equiv \left|\mathcal{N}_i^B\right| = 0.002 \times |B|$. These matrices can be used to transfer quantities across the two datasets. For instance, the inferred mRNA expression level of gene m in the CODEX cell i is given by

$$S_{jm}^{CODEX} = \sum_i \mathcal{M}_{ij}^{CITEseq\to CODEX} S_{im}^{CITEseq}$$

where $S^{CITEseq}$ denotes the mRNA expression matrix in the CITE-seq dataset. Similarly, the mRNA cell populations can be mapped to the CODEX data using

$$C_{jm}^{CODEX} = \sum_i \mathcal{M}_{ij}^{CITEseq\to CODEX} C_{im}^{CITEseq}$$

where the sum runs over all cells in the CITE-seq dataset and $C_{ic}^{CITEseq}$ is the indicator function of cluster c. Note that due to the mapping uncertainties, the resulting feature vector is no longer a binary vector. To assess mapping uncertainties (Fig. 2c), we computed the Pearson's correlation coefficient of the vectors C_{jc}^{CODEX} that result from restricting the above sum to cells in each of the two mice profiled with CITE-seq.

Parameter selection

For different values of k_{anchor}, k_{filter}, and k_{score}, we evaluated the performance of the algorithm to accurately map a set of "gold standard" cell populations. The populations we considered were B cells, T cells, NK cells, dendritic cells, neutrophils, plasma cells, and red-pulp macrophages, as they were general enough to be clearly identifiable in both datasets by clustering and the expression of specific markers. We used the Louvain community detection algorithm in a $k = 49$

nearest neighbor graph for clustering the CODEX protein data. To quantify the performance of the mapping, we defined the quality scores of a set of anchors $\mathcal{A}$ as

$$Q_{\mathcal{A}}^{u} = |\mathcal{A}|_{u}^{-1} \sum_{(i,j)\in\mathcal{A}} z_{(i,j)}^{u}$$

$$z_{(i,j)}^{\text{anchor}} \equiv z_{(i,j)}^{\text{filter}} \equiv \begin{cases} 1, & c_i = c_j \in \mathcal{C} \\ 0, & \text{otherwise} \end{cases}, \qquad |\mathcal{A}|_{\text{anchor}} \equiv |\mathcal{A}|_{\text{filter}} \equiv |\mathcal{A}|$$

$$z_{(i,j)}^{\text{score}} \equiv \begin{cases} s_{ij}, & c_i = c_j \in \mathcal{C} \\ 0, & \text{otherwise} \end{cases}, \qquad |\mathcal{A}|_{\text{score}} \equiv \sum_{(i,j)\in\mathcal{A}} s_{ij}$$

where c_i is the cell type of cell i and $\mathcal{C}$ is the set of "gold standard" populations. We sequentially chose the values of k_{anchor}, k_{filter}, and k_{score} that maximized these quality scores.

Spatial relationship among cell populations

To assess the spatial relationship between two feature vectors f and g defined over the cells in the CODEX dataset, we built a $k = 2$ nearest neighbor graph using Euclidean distance in the CODEX spatial dimensions expressed in nm. We then introduced the adjacency score, defined as

$$D(f, g) = f^{T} \cdot A \cdot g$$

where A is the adjacency matrix of the nearest neighbor graph. This score takes high values when the features take high values in adjacent cells. The scale of the interactions is set by the magnitude of the nearest neighbor parameter k. Features that we have used in this paper include cell population assignments C_{jc}^{CODEX} (to assess whether two cell populations co-localize spatially) and mapped gene expression S_{jm}^{CODEX} (to assess whether genes encoding for ligands and receptors are expressed in adjacent cells). The significance of this score was assessed using a null distribution built by permuting the cell ID's. For mutually exclusive binary features (such as cluster assignments) the null distribution can be computed analytically in terms of the hypergeometric distribution $\text{Hypergeom}(u; N, K, n)$,

$$Prob(D(f, g) = u) \sim \text{Hypergeom}(u; v(v-1), 2(f^{T} \cdot I)(g^{T} \cdot I), m)$$

where v and m are the number of nodes and edges in the nearest neighbor graph, respectively, and I is the identity matrix. For non-binary features, we did not find a closed form for the null distribution, so we approximated it using a normal distribution whose parameters were estimated from 1,000 random permutations. We controlled the false discovery rate for multiple hypothesis testing using the Benjamini-Hochberg *q*-value procedure.

To account for the effect of mapping uncertainties on the adjacency score of cell populations, we also computed the overlap score, $f^T \cdot g$, and assessed its significance by randomly permuting the entries of one of the feature vectors. In addition, we evaluated the Pearson's correlation of the adjacency score *q*-values across the three mice profiled with CODEX (Fig. 3b).

Identification of paracrine interactions

We used CellPhoneDB[23] to identify significant ligand-receptor pairs within the CITE-seq mRNA expression data. CellPhoneDB identifies genes encoding for ligand and receptor pairs that are differentially expressed in one or more cell populations using a curated database of ligands and receptors. Since CellPhoneDB only considers human gene pairs, we generated a mouse ortholog database of ligands and receptors using Ensembl[40] (version 96). For simplicity, this analysis was restricted to only those genes which have a unique ortholog. The results of this analysis were then filtered using the adjacency score approach described above (see "Spatial relationship among cell populations") to identify pairs of genes significantly expressed in adjacent cells. The expression of any complexes output by CellPhoneDB was calculated as the sum of the expression of their component genes.

Online database

The complete results of our analysis can be interactively queried through a web application hosted at the URL: https://camara-lab.shinyapps.io/stvea .

STvEA software

All algorithms have been implemented and documented in an R package. The package can be downloaded from the URL: https://github.com/CamaraLab/STvEA .

Acknowledgements

The authors would like to thank Daniel Aldea, Rachael Aubin, Jake Crawford, Doug Epstein, Yugong Ho, Jesse Humenik, Yana Kamberov, Steve Liebhaber, Fernanda Mafra, Michael Peel, Renata Pellegrino, Qi Qiu, Staci Rakowiecki, Jacqueline Smiler, Ryan Staupe, John Wherry, Hao Wu, and the Zhou lab for scientific discussions and assistance with various aspects of the experiments. The work of P.G.C. and S.W. is partially funded by Stand-Up-To-Cancer Convergence 2.0.

Author Contributions

E.C.T. performed all the experiments. K.W.G. implemented all the algorithms and performed the analyses. S.W. designed and implemented the web application. P.G.C. conceived the study and supervised the work. All authors contributed to writing the manuscript.

Competing Interests

The authors declare no competing interests.

Figures

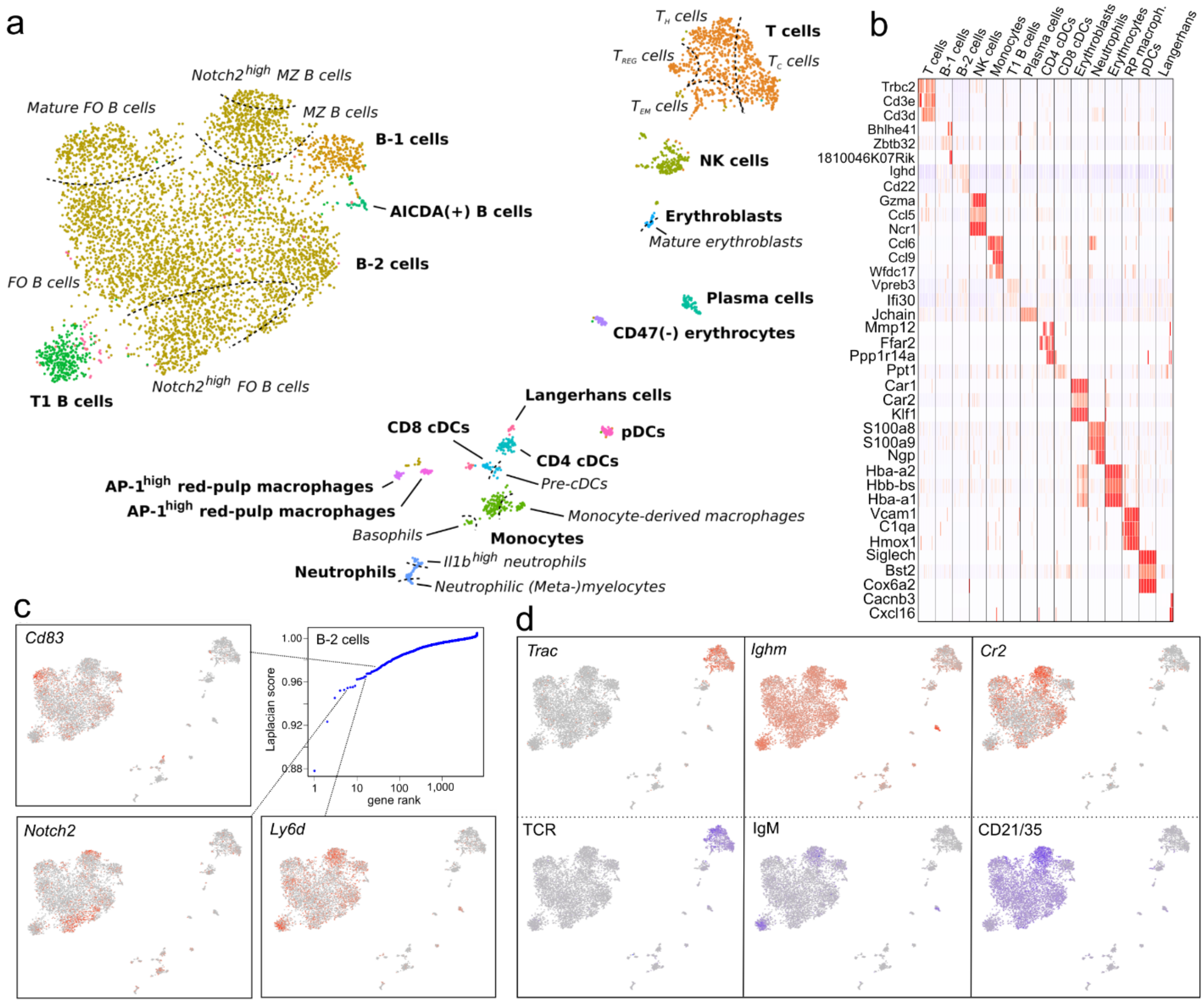

Figure 1. A high-resolution CITE-seq atlas of the murine spleen. a) UMAP representation of the mRNA expression data of 7,097 cells from the murine spleen profiled with CITE-seq. Cell populations were identified by clustering (represented in different colors) and annotated by differential expression analysis (bold text) and a spectral graph method (italic text; see also panel c). **b)** Heatmap depicting the expression of some of the top differentially expressed genes in each cluster. **c**) Analysis of the cellular heterogeneity within the clusters of B-2 cells using a spectral graph approach. Genes were ranked according to their Laplacian score and the statistical significance was assessed for each gene by randomization. In the figure, the expression levels of some of the significant genes is depicted in the UMAP representation. The complete results are provided for all clusters in Supplementary Data 4. **d)** mRNA expression levels of *Cr2*, *Ighm*, and *Trac* (top) and the expression levels of the proteins they code for (bottom).

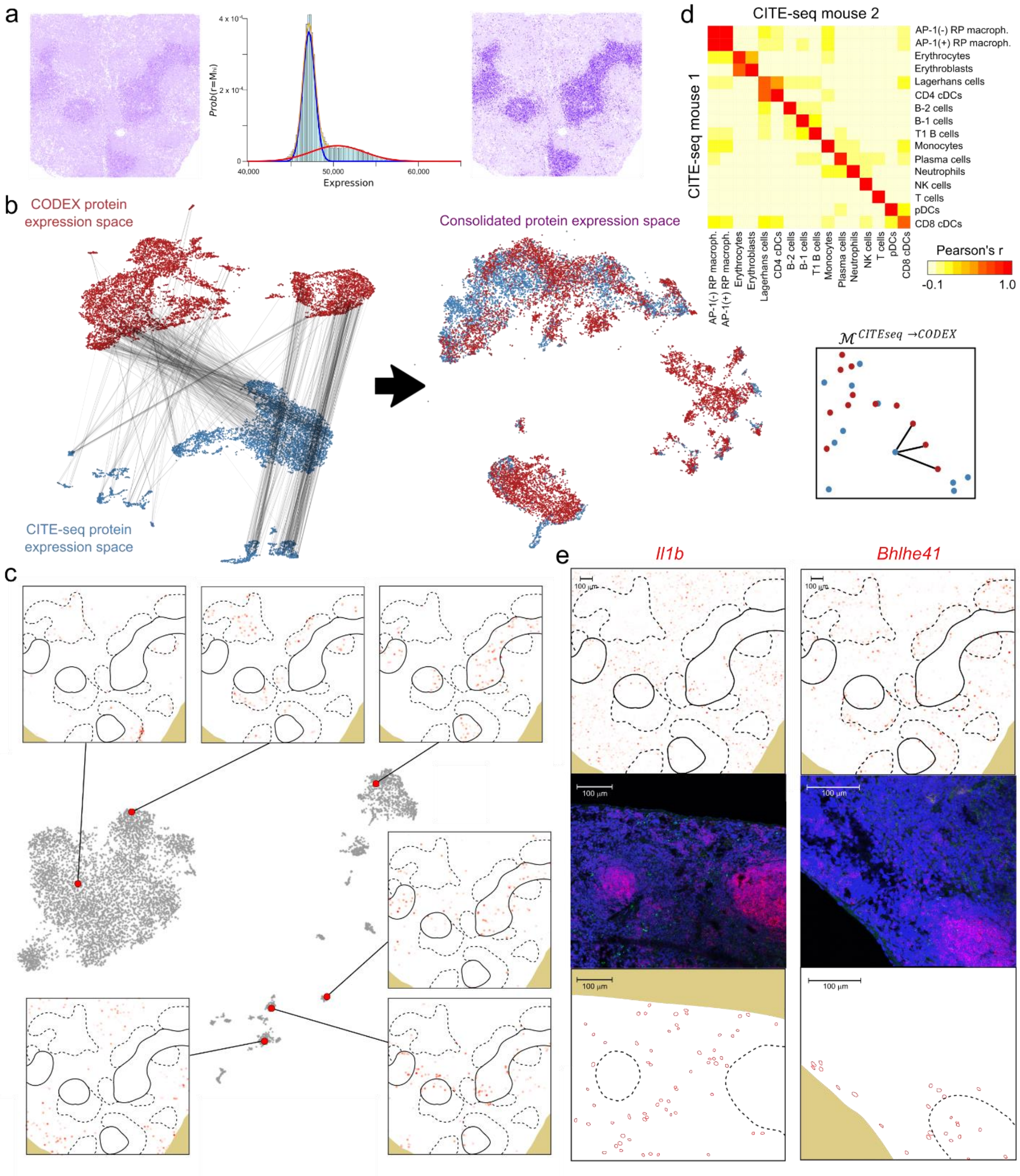

Figure 2. Mapping of the splenic CITE-seq atlas into histology sections profiled with CODEX. a) Normalization of CODEX data using a Gaussian mixture model. The levels of CD4 protein determined by CODEX are shown in a splenic section after standard processing of the data (left). A two-component Gaussian mixture model was fit to the CD4 protein levels, where the lowest and highest mode components correspond to background and signal, respectively

(middle). Upon filtering out the background component, the CD4 signal at T-cell follicles becomes more evident. **b)** Schematics of the procedure to map the CODEX and CITE-seq protein expression spaces. A set of anchors are identified using a mutual nearest neighbors approach and weighted according to their degree of consistency with the mRNA expression space (left). The anchors are used to merge the CODEX and CITE-seq protein expression spaces into a common space (middle). The transfer matrix $\mathcal{M}^{CITEseq \to CODEX}$ is built by looking at the nearest CODEX cells to each CITE-seq cell in the merged protein expression space (right). **c)** Mapping of individual cells from the splenic CITE-seq atlas into a murine splenic section profiled with CODEX. The figure shows the inferred locations of 6 cells from the CITE-seq atlas in the splenic section. For reference, the T and B cell zones in the tissue section are indicated with continue and dashed lines, respectively. As it can be seen in the figure, transcriptomic differences between cells of the same cell type often correspond to different spatial locations. These differences were consistent when the same cells were mapped to other murine splenic sections profiled with CODEX (Supplementary Fig. 3). **d)** Consistency between the mappings of two different murine spleens profiled by CITE-seq into the same CODEX dataset. A heatmap showing the correlation between the CODEX cell assignments for each cell population in the two mice profiled with CITE-seq. In an ideal scenario, diagonal entries would be perfectly correlated (Pearson's $r = 1$) and off-diagonal entries anti-correlated (Pearson's $r < 0$). Departures from that situation quantify mapping inaccuracies. As represented in the figure, STvEA has an excellent performance for most splenic cell populations, with the most notable inaccuracies being between AP-1^{high} and AP-1^{low} red-pulp macrophages, and erythrocytes and erythroblasts. **e)** mRNA expression levels predicted by STvEA (top) and measured by RNA FISH (middle and bottom) in murine splenic sections for the genes *Il1b* (left) and *Bhlhe41* (right). Red: *Cd79a*, green: *Il1b* / *Bhlhe41*, blue: DAPI. T and B cell zones in the tissue sections are indicated with continue and dashed lines, respectively. To facilitate interpretation, the relative location of cells expressing *Il1b* and *Bhlhe41* is indicated at the bottom.

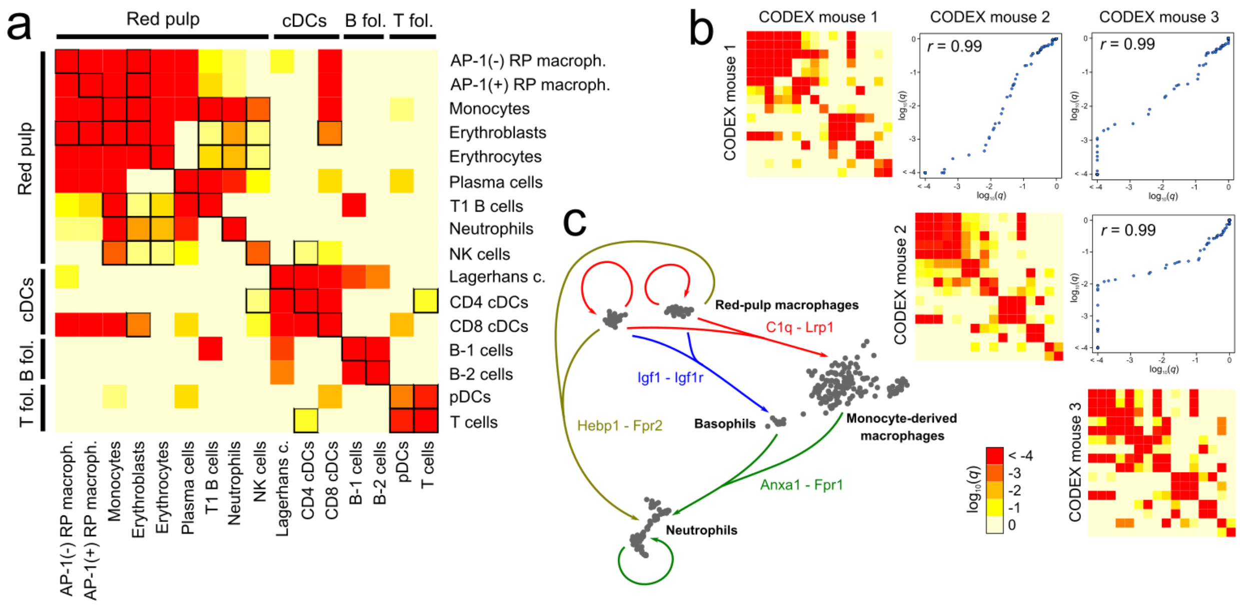

Figure 3. Identification of cell-to-cell interactions among splenic cell populations. a) Heatmap showing the significance of the spatial co-localization of splenic cell populations, inferred by STvEA. Significant relations (*q*-value ≤ 0.05) that cannot be explained by mapping errors (95% CL) are indicated with black squares. **b)** Consistency of the analysis across different splenic sections profiled by CODEX. Three different splenic sections were independently mapped to the CITE-seq atlas. The inferred spatial co-localization patterns are highly consistent across the three sections, as indicated by the high Pearson's correlation coefficients among the estimated significances. **c)** Some of the significant potential paracrine interactions among red-pulp macrophages, basophils, neutrophils, and monocyte-derived macrophages in the red pulp. Interactions were inferred based on the differential expression of the genes encoding for the ligand and receptor and on their spatial co-localization.

References

1. Stack, E.C., Wang, C., Roman, K.A. & Hoyt, C.C. Multiplexed immunohistochemistry, imaging, and quantitation: a review, with an assessment of Tyramide signal amplification, multispectral imaging and multiplex analysis. *Methods* **70**, 46-58 (2014).
2. Goltsev, Y. et al. Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell* **174**, 968-981 e915 (2018).
3. Jungmann, R. et al. Multiplexed 3D cellular super-resolution imaging with DNA-PAINT and Exchange-PAINT. *Nat Methods* **11**, 313-318 (2014).
4. Cornett, D.S., Reyzer, M.L., Chaurand, P. & Caprioli, R.M. MALDI imaging mass spectrometry: molecular snapshots of biochemical systems. *Nat Methods* **4**, 828-833 (2007).
5. Giesen, C. et al. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat Methods* **11**, 417-422 (2014).
6. Angelo, M. et al. Multiplexed ion beam imaging of human breast tumors. *Nat Med* **20**, 436-442 (2014).
7. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* **14**, 865-868 (2017).
8. Peterson, V.M. et al. Multiplexed quantification of proteins and transcripts in single cells. *Nat Biotechnol* (2017).
9. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat Methods* **15**, 1053-1058 (2018).
10. Govek, K.W., Yamajala, V.S. & Camara, P.G. Spectral Simplicial Theory for Feature Selection and Applications to Genomics. *arXiv preprint arXiv:1811.03377* (2018).
11. He, X., Cai, D. & Niyogi, P. in Advances in neural information processing systems 507-514 (2006).
12. Jaitin, D.A. et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776-779 (2014).
13. Tabula Muris, C. et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367-372 (2018).
14. Han, X. et al. Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172**, 1091-1107 e1017 (2018).
15. Lewis, S.M., Williams, A. & Eisenbarth, S.C. Structure and function of the immune system in the spleen. *Sci Immunol* **4** (2019).
16. Bronte, V. & Pittet, M.J. The spleen in local and systemic regulation of immunity. *Immunity* **39**, 806-818 (2013).
17. Stuart, T. et al. Comprehensive Integration of Single-Cell Data. *Cell* (2019).
18. Haghverdi, L., Lun, A.T.L., Morgan, M.D. & Marioni, J.C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* **36**, 421-427 (2018).
19. Eisenbarth, S.C. Dendritic cell subsets in T cell programming: location dictates function. *Nat Rev Immunol* **19**, 89-103 (2019).
20. Wang, F. et al. RNAscope: a novel in situ RNA analysis platform for formalin-fixed, paraffin-embedded tissues. *J Mol Diagn* **14**, 22-29 (2012).
21. Kreslavsky, T. et al. Essential role for the transcription factor Bhlhe41 in regulating the development, self-renewal and BCR repertoire of B-1a cells. *Nat Immunol* **18**, 442-455 (2017).
22. Wen, L., Shinton, S.A., Hardy, R.R. & Hayakawa, K. Association of B-1 B cells with follicular dendritic cells in spleen. *J Immunol* **174**, 6918-6926 (2005).
23. Vento-Tormo, R. et al. Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature* **563**, 347-353 (2018).

24. Ogden, C.A. et al. C1q and mannose binding lectin engagement of cell surface calreticulin and CD91 initiates macropinocytosis and uptake of apoptotic cells. *J Exp Med* **194**, 781-795 (2001).
25. Gao, J.L. et al. F2L, a peptide derived from heme-binding protein, chemoattracts mouse neutrophils by specifically activating Fpr2, the low-affinity N-formylpeptide receptor. *J Immunol* **178**, 1450-1456 (2007).
26. Sugimoto, M.A., Vago, J.P., Teixeira, M.M. & Sousa, L.P. Annexin A1 and the Resolution of Inflammation: Modulation of Neutrophil Recruitment, Apoptosis, and Clearance. *J Immunol Res* **2016**, 8239258 (2016).
27. Hartnell, A. et al. Identification of selective basophil chemoattractants in human nasal polyps as insulin-like growth factor-1 and insulin-like growth factor-2. *J Immunol* **173**, 6448-6457 (2004).
28. Hirai, K. et al. Modulation of human basophil histamine release by insulin-like growth factors. *J Immunol* **150**, 1503-1508 (1993).
29. Koketsu, R. et al. Activation of basophils by stem cell factor: comparison with insulin-like growth factor-I. *J Investig Allergol Clin Immunol* **18**, 293-299 (2008).
30. Ochensberger, B., Daepp, G.C., Rihs, S. & Dahinden, C.A. Human blood basophils produce interleukin-13 in response to IgE-receptor-dependent and -independent activation. *Blood* **88**, 3028-3037 (1996).
31. Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S. & Zhuang, X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
32. Moffitt, J.R. et al. High-throughput single-cell gene-expression profiling with multiplexed error-robust fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 11046-11051 (2016).
33. Lee, J.H. et al. Highly multiplexed subcellular RNA sequencing in situ. *Science* **343**, 1360-1363 (2014).
34. Eng, C.L. et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568**, 235-239 (2019).
35. Lubeck, E., Coskun, A.F., Zhiyentayev, T., Ahmad, M. & Cai, L. Single-cell in situ RNA profiling by sequential hybridization. *Nat Methods* **11**, 360-361 (2014).
36. Codeluppi, S. et al. Spatial organization of the somatosensory cortex revealed by osmFISH. *Nat Methods* **15**, 932-935 (2018).
37. Rodriques, S.G. et al. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463-1467 (2019).
38. Vickovic, S. et al. High-density spatial transcriptomics arrays for in situ tissue profiling. *bioRxiv*, 563338 (2019).
39. Merritt, C.R. et al. High multiplex, digital spatial profiling of proteins and RNA in fixed tissue using genomic detection methods. *BioRxiv*, 559021 (2019).
40. Zerbino, D.R. et al. Ensembl 2018. *Nucleic acids research* **46**, D754-D761 (2018).