# SPEX: A modular end-to-end analytics tool for spatially resolved omics of tissues

Ximo Pechuan-Jorge*,Xiao Li*,Tyler Risom, Artem Zubkov, Evgeniy Tabatsky, Alexander Prilipko, Xin Ye, Zhen Shi, Malgorzata Nowicka, Frank Peale, Derrek Hibar, James Ziai, Raj Jesudason†, Darya Orlova†

[1] Genentech, Inc., 1 DNA Way, South San Francisco, CA 94080
∗ Denotes equal contribution. † To whom correspondence should be addressed: jesudason.rajiv@gene.com and orlovad@gene.com

**Recent advancements in transcriptomics and proteomics have opened the possibility for spatially resolved molecular characterization of tissue architecture with the promise of enabling a deeper understanding of tissue biology in either homeostasis or disease. The wealth of data generated by these technologies has recently driven the development of computational pipelines that, nevertheless, have the requirement of coding fluency to be applied. To remove this hurdle, we present SPEX (Spatial Expression Explorer), a comprehensive image analysis software implemented as a user-friendly web-based application with modules that can be put together by the user as pipelines conveniently through a graphical user interface. SPEX's infrastructure allows for streamlined access to open source image data management systems and analysis modules for cell segmentation, cell phenotyping, cell-cell co-occurrence and spatially informed omics analyses. We demonstrate SPEX's ability to facilitate the discovery of biological insights in spatially resolved omics datasets from healthy tissue to tumor samples.**

# 1 Introduction

Spatially resolved, highly multiplexed protein and RNA profiling of tissues at the cellular or subcellular level is essential for the understanding the molecular components underlying tissue architecture and function (Dries et al., 2021). This fine characterization enables the measurement of cellular heterogeneity and cell-cell interactions in the spatial dimension which in turn brings us closer to identify molecular determinants of tissue function. Hence, recently developed technologies, like IMC (Giesen et al., 2014), MIBI (Keren et al., 2018),CODEX (Goltsev et al., 2018) or MERFISH (Chen et al., 2015), that allow for the acquisition of spatially resolved proteomic and transcriptomic states of individual cells and tissues, hold a great promise for advancing our understanding in many areas of cell and developmental biology. However, the widespread adoption of these technologies vitally depends on the development of analytical capabilities and infrastructure to readily extract biologically relevant conclusions from the complex data generated.

Spatially resolved omics data contain multiple layers of information which in turn are derived from imaging data and the molecular measurements associated with it. Examples of said layers constitute cell morphology, patterns of protein and transcript expression, cell neighborhoods or cell-cell communication at different spatial scales. To obtain this information, robust pre-processing and visualization methods are required as well as ways of analyzing the omics information and integrating it with the image derived properties (Palla et al., 2022a). The current software ecosystem developed in the past five years leverages about ten common analysis categories, including cell segmentation, clustering and spatial analytics, all required for meaningful interpretation of spatially resolved omics data (Lewis et al., 2021). Nevertheless, currently existing tools and custom analysis protocols provide only fragmentary solutions (Martinelli and Rapsomaniki, 2022; Palla et al., 2022b). None of them by itself provide means for efficient data representation and pre-processing, comprehensive end-to-end analysis pipeline and interactive visualizations. Thus, to harness the power of multiplexed spatial data, researchers are forced to push their data through multiple standalone data visualization and analytical tools that were not designed to fit into one pipeline and often have some unique prerequisites for the input data format. Hence, current exploration of spatially resolved omics data is a non-scalable process that requires coding abilities.

To remove this stumbling block, we developed SPEX (Spatial Expression Explorer), a modular end-to-end analytics tool. SPEX is an application that could readily be deployed on cloud, or on a local virtual machine. It has a unique architecture that allows users to build data modality-specific customized pipelines from the SPEX's adopted/novel cutting edge analytics developments or via plugging in external analysis modules. To support the needs of the cross-functional research community we designed SPEX in a way that one could operate it either from the user interface (code-free environment) or via more involved hands-on coding.

Here, as part of the SPEX software package, we also introduce new methods to do per-sample immunophenotyping, cell community detection and quantitative assessment of cell-cell co-occurrence. For immunophenotyping, we rely on per-sample clustering and further cluster alignment between samples that utilizes distance metric learning with UMAP (McInnes et al., 2018) and cluster matching with QFMatch (Orlova et al., 2018). To quantitatively assess the local and global patterns of spatial association between the two cell categories here we apply the Colocation Quotient (CLQ) score common in other spatial analysis domains. For community detection, we determine the matrix of neighborhood compositions at a certain radius and cluster them to identify the cell communities. To aid in the biological interpretation of the data we integrate known pathway analysis and cell-cell communication inference methodologies with the spatial information provided by our spatial

analytics and the neighborhood composition.

Collectively, as we show here, these analytical approaches helped us to identify new patterns in tumor-immune microenvironment composition in a set of heterogeneous samples from a cohort of triple negative breast cancer patients. In another example, we applied SPEX community pathway analysis to human lung cancer specimen and elucidated the pathways associated with immune attraction/avoidance in immune cells and tumor cells.

# 2 Results

## 2.1 SPEX General description and architecture

SPEX is a comprehensive image analysis pipeline implemented as a user-friendly web-based application with flexible plug-in analysis modules (Figure 1, A). SPEX provides both infrastructure and quantitative analysis methods that allow for efficient store, manipulation and interactive visualization of spatial expression data. SPEX has an easy to use graphical user interface which allows the wider research community to build analytical pipelines and visualize high dimensional spatial data.

The analysis modules provide the end-to-end analysis by implementing a wide range of algorithms for cell segmentation, cell phenotyping, spatial expression patterns, and cell-cell cooccurrence. These modules are flexible and self-contained depending on inputs from upstream analysis modules and generating outputs that in turn feed downstream analysis modules. This means that they can be put together as pipelines defined by the user through the graphical user interface.

SPEX's architecture design (Figure 1, B) is mainly driven by the requirements to the system performance, availability, and scalability. In particular, we aimed for an architecture supporting multi-user access to computing resources while ensuring and guaranteeing the execution of the resource-heavy tasks set by the user in a reasonable time. As an additional requirement, we considered the ease of system extension and maintenance. All these requirements are fulfilled by the microservice architecture that is inherently flexible as it is based on the idea of interoperability of small, loosely coupled and easily modifiable modules (microservices). A microservice is a small program that performs a clearly assigned task. Each microservice is hosted in a Docker container and can run individually or as a group via docker-compose. SPEX microservices communicate through the data bus instrumented by Redis. Redis manages message transfer between the services ensuring messages order and guaranteed delivery even if temporary system outages occur. Message distribution and load balancing mechanisms allow to easily increase the number of the microservices' instances and spread them across different servers enabling straightforward scaling and performance boosting possibilities.

## 2.2 Image Processing Module

SPEX includes a modular pipeline to facilitate tissue-based single-cell segmentation. This generalized pipeline aims to accommodate a wide range of high-dimensional imaging modalities such as IMC, MIBI, cyclic IF and MERFISH.

Image processing is executed in a 5 step sequence with the ultimate goal of generating a cell by expression matrix. These steps include; image loading, image preprocessing, single-cell segmentation, post-processing and feature extraction (Figure 2,A). Each step contains a selection of modules which can be linked together to address the particularities of a given image set (Figure S2).

The adaptability of the pipeline was demonstrated by analyzing imaging data coming from a variety of high dimensional imaging modalities which included IMC, MIBI and MERFISH. Each imageset presented unique characteristics; varying in resolution, signal to noise and structural attributes (Figure 2,B).

To demonstrate the module capabilities, a publicly available Multiplex Ion Beam Imaging (MIBI) dataset was processed through the SPEX image processing module. This imageset consisted of 41 field of view images with 36 channels. Being a non-optical mass detection platform, it is common for raw images to include substantial noise when compared to traditional optical acquisition platforms. Therefore, median denoising was executed on channels utilized for segmentation (dsDNA, H3K9ac, H3K27me3). These channels were then merged and then segmented for single-cells using the Mesmer deep learning model. We found strong concordance with the previously identified phenotype compositions thereby underscoring the flexibility of the pipeline in recapitulating prior image processing methodologies (Figure 2,C).

Similar pipelines were generated for IMC and MERFISH with the difference that Stardist was used for single-cell segmentation. With the added parameterization of upsampling or downsampling the pixel data to better match the training data used by the model. Comparatively, IMC is a low resolution acquisition with a native resolution of 1 $\mu m$/pixel while MERFISH image data was 105 $nm$/pixel. We found that resizing pixels with interpolation to a scale more akin to a 20x virtual magnification facilitated high quality segmentation by StarDist (Schmidt et al., 2018).

To assess sensitivity of phenotyping readouts to segmentation model selection, we compared segmentation of the publicly available TNBC MIBI imaging data segmented with either StarDist or Mesmer then clustered with FLOWSOM in R (external from SPEX) . We show that ranked global population levels of phenotypes share the same trend, with epithelial cells and macrophages representing the majority of cells in the dataset. The interchangeable nature of the SPEX modules illustrates task-specific flexibility. End users are encouraged to select modules which best fit the needs of their particular experiment and imaging data.

## 2.3 Clustering Module

To enable the analysis of the omics data associated with the images, SPEX provides a clustering module that accommodates both proteomics and transcriptomics inputs. For transcriptomics, we rely on the suite of proven methods offered by the Pegasus package (Li et al., 2022). For proteomics, we also provide a set of clustering algorithms to enable cell type identification that are common in the field such as Phenograph (Levine et al., 2015).

The previously mentioned set of methods can be applied to a single sample or multiple samples, nevertheless as in their droplet based counterparts, when considering multiple samples, spatially resolved omics can often also require an integration step prior to downstream analysis. Integration removes potential batch effects and ensures that cells across samples are given a comparable representation that allows for the identification of cell types that are robust and biologically meaningful. SPEX provides a novel proteomics integration method that makes use of an annotated reference set to train an embedding where the test samples are projected (Figure 3, A).

We used this integration method to project 41 MIBI samples from a triple negative breast cancer cohort (Keren et al., 2018) into the same embedding space to discern the different tumor microenvironment compositions. The joint embedding allowed for the identification of two tumor cell clusters in particular sample (Figure 3, B). When inspected in further detail, one of the clusters exhibited a higher stromal signature that could be due to differentiation of the tumor lineage or

because its adjacency to stroma in the tissue sample (Figure 3, C). Finally, we demonstrated the utility of this clustering algorithm by determining the cell type composition of the entire cohort which appears to be consistent to what is reported on (Keren et al., 2018) (Figure 3, D).

## 2.4 Spatial Analyses Module

Tissue architecture can be complex, with cell types forming spatial patterns that define particular domains where they might exert very different functions owing to the distinct local cellular context. These functions might be reflected in the associated gene or protein expression patterns. SPEX implements two modules that integrate the omic and the spatial component of the data at different levels of spatial resolution. Both modules label the cells according to their spatial properties which is then utilized to perform differential expression analysis, pathway analyses and cell-cell communication analysis.

The first module relies on the CLQ measurements to stratify the data in terms of co-location or avoidance between cell types. The second module explores the expression patterns of locally defined cell communities. We explored both modules in a human lung cancer sample stained with a MERFISH panel of 484 genes. Using cell type specific marker genes, we first determined the major cell types including immune, stromal, endothelial, and tumor cells to feed into the subsequent spatial analysis.

### 2.4.1 Spatial Analytics Enable Tumor Immune Phenotyping

Tumor immune phenotypes have become an important clinical classification with actionable therapeutic consequences for patients (Chen and Mellman, 2017; Galon and Bruni, 2019). In relationship with their immune contexture, tumors are classified as immune inflamed if T cells have infiltrated the tumor nests, and as immune excluded if T cells are preferentially located in the stroma. If few-to-no T cells are to be found in the tumor or its adjacent stroma, the tumor is classified as an immune desert.

We illustrate the SPEX spatial analysis module using the CLQ method to characterize the immune phenotypes of a published breast cancer cohort (Keren et al., 2018). Briefly, the CLQ measures the co-occurrence or avoidance of cell type pairs by looking at the local density of a target cell type at a fixed radius from each cell of the sample belonging to a reference cell type (Figure 4, A). That way, one can, for instance, query if target $CD8^+$ T cells are co-locating or avoiding tumor cells which intuitively corresponds to the inflamed and excluded tumor immune phenotypes. Because the CLQ method relies on a continuous summary statistic, it has the ability to go beyond the three categories commonly accepted and highlight potentially cases that lie at the boundary of these classes or identify novel tumor immune spatial phenotypes.

To determine the tumor immune phenotypes of the breast cancer cohort, we adopted a two-step approach relying on a Gaussian mixture model (see Methods for details). Based on a hard-cutoff threshold, we identified 7 samples which have less than 200 immune cells as belonging to the desert immune phenotype. For the remaining 34 samples, the Gaussian mixture model rendered two clusters, with 9 samples classified as immune inflamed, having smaller mean values of $CLQ_{tcells \rightarrow tumor}$ and $CLQ_{tumor \rightarrow tcells}$, and 25 samples classified as immune excluded, where the CLQ-calculated spatial enrichment between tumor and immune cells is reduced (Figure 4, B).

Having defined the tumor immune phenotypes with CLQ analysis as desert, excluded, or inflamed tumors, we explored potential gene expression patterns yielded by this classification. Of clinical relevance, we observed a significant increase in PDL1 expression by tumor-associated

macrophages in inflamed tumors suggesting the tumor-immune spatial relationships may represent informative biomarkers of immunotherapy potential (Figure 4, C) (Johnson et al., 2018; Tsakiroglou et al., 2020). We leveraged CLQ analysis to further interrogate macrophage phenotypes in these tumors by computing their pairwise spatial enrichment with all other cell types. Interestingly, this combination of CLQ and expression analysis revealed that macrophages that are spatially enriched (attractive) with CD8+ Cytotoxic T Cells had significantly higher IDO1 expression compared to macrophages that are spatially isolated (avoidant) from these T Cells (Figure 4, D). This relationship was observed in excluded tumors and further enriched in inflamed tumors, suggesting that in addition to an enhanced potential for PDL1 antagonists in this inflamed subset of TNBC patients, IDO1 may be an immunosuppressive target of interest to alleviate local macrophage suppression of T cell activity (Liu et al., 2009).

#### 2.4.2 Spatial Analytics Integration with Omics Data

The CLQ module provides an effective summary statistic to capture relations of spatial avoidance or attraction between pairs of cell types in an imaged tissue. Utilizing the CLQ to inform omics analysis can allow the identification of genes and pathways with an expression state that is correlated with said pairwise spatial relationships. CLQ thus provides insights on how such co-occurrence alters the transcriptome of the cell type pair of interest.

As an example, we used the CLQ method to stratify a human lung cancer MERFISH sample by the relationship of attraction or avoidance for each cell type pair (Figure 5, A). We investigated how proximity to T cells is accompanied by inflammatory and cell proliferation pathways. We noticed up-regulation of RTK/MAPK pathway genes and down regulation of Trail in cells proximal to T cells across epithelial, myeloid and endothelial cells. Interestingly, this expression pattern relationship is inverted for B cells. Furthermore, the integration of the CLQ with omics data allowed us to identify genes that are differentially expressed in the proximal presence or absence of another cell type. As an example, we assessed how gene expression differs between T cells that are classified as "attractive" to or "avoidant" of Tumor cells (as determined by CLQ analysis). We identified several significant differences, however we want to caution that the problem of freely diffusing RNA or ambient RNA on MERFISH data can impact the relevance of these observations. For example, the top differentially expressed gene was SFTPB, a surfactant protein essential for lung function which is probably associated with the fact that said T cells with such expression are proximal to the lung epithelium. Optimizing segmentation and further de-noising the sample with spatial aware algorithms is important to circumvent this caveat.

The CLQ method is limited to pairwise relations and does not take into account higher order interactions with additional cell types within potential functional communities composed of multiple cell types. We acknowledge that spatial patterns exist beyond pairwise spatial correlations and as such we introduce the next analysis module, SPEX community analysis. This module identifies cell communities by determining the neighborhood composition for each cell in the image at a given radius and then clustering all the different neighborhoods to identify cell communities.

In the lung tumor MERFISH sample, the resulting analysis demonstrated the presence of cell type communities that define regions of interest in the image with potentially different biology. This allowed us to extract further information than the mere projection of the cell types identified by transcriptomics back into the image (Figure S3, A). SPEX communities, when highlighted in the tissue slide allow for the identification of potential structures. For example, SPEX communities bring out aggregates of B cells surrounded by layer comprised of a combination of B cells, epithelial

cells and T cells potentially defining tertiary lymphoid structures (Figure 5, B). Beyond the spatial information, the clusters themselves can be characterized using pathway analysis. Pathways will be present as a function of the cell types involved in the community and their contextual transcriptomic state, In the present case, TGF$\beta$ has larger activity in communities containing more fibroblasts (Figure S3, C).

The SPEX community analysis module can also be leveraged to look at specific cell types in the context of a particular microenvironment. More concretely, differential gene expression analyses for the same cell type across clusters can further identify context dependent expression patterns (Figure 5, C). We highlight how the T cells tend to be expressing more CCR7 in SPEX community 5 compared to 1 potentially indicating a differentiation towards more mature T cells in that community. Tumor cells also show some interesting patterns, when associated with stroma they express more HLA and CYR61, involved in angiogenesis. Other cell types have interesting contextual gene expression patterns (Figure S3, D) like endothelial cells exhibiting more proliferation markers in the context of a rich immune microenvironment relative to when adjacent to serous cells potentially reflecting tissue homeostasis. Overall, our community module generates hypotheses about regions of the tissue where potential contextual phenotypes are observed and guides the further exploration of the data.

## 2.5 SPEX Pipeline Recapitulates Tonsil Tissue Composition

In the previous sections, we illustrated the capabilities of the different modules that integrate SPEX. To demonstrate that SPEX provides an end-to-end analysis solution, we validated the SPEX workflow by analyzing a sample from human tonsil tissue, which has a well-defined cellular organization (Allen et al., 2007; Bannard and Cyster, 2017; Willard-Mack, 2006). To achieve this, a 4$\mu$m-thick section of formalin-fixed, paraffin-embedded (FFPE) tonsil tissue was stained with a panel of Imaging Mass Cytometry (IMC) antibodies listed in Supplementary Table 1. The antibody panel was designed to simultaneously characterize the composition of the immune compartment, the spatial relationship between immune cells and stromal cells, and the interactions among cell subsets.

The stained tonsil sample was imaged with the Hyperion Imaging System and then pre-processed and analyzed using a SPEX pipeline utilizing most of SPEX modules (Figure 6, A). This pipeline applied to the tonsil tissue IMC images started with per cell segmentation using StarDist based on the nuclear histone H3 channel. Then, Phenograph was used for clustering the segmented cells according to their median marker expression levels of the different IMC channels. Clustering outcomes were visualized using UMAP, and cluster labels were assigned based on the IMC marker measurements as well as positional data (x and y coordinates) of cells that comprise identified clusters (Figure 6, B-E).

The resulting UMAP clustering generated 3 well-defined lobes representing epithelial, follicular and paracortical tissue compartments distinguished primarily by expression of cytokeratin, Ki67, and CD8, respectively, consistent with the composition of cells in these compartments. Within the individual UMAP lobes, the location of closely related cell populations correlates with biologically relevant parameters, for example, the graded expression of Ki67 and BCL2 in maturing germinal center B-cells (Figure 6, C,D).

The CLQ analysis, from the SPEX Spatial Analysis module, appropriately highlights spatial colocalization or separation between observable populations (Figure 6, F,G). We further systematically assessed pairwise patterns of non-random co-occurrence or avoidance of the identified cell

types by using a permutation test to compare the number of interactions between all cell types (in a user defined neighborhood) in a given image to that of synthetic matched controls generated under the null hypothesis, that is containing randomized cell phenotypes. The neighborhood radius (20 $\mu$m) was chosen empirically to restrict cell-cell co-occurrence assessment to only those cells that are close enough to potentially establish physical contact between their membranes. This approach, similar to the one described in the histoCAT paper (Schapiro et al., 2017), allows to determine the significance of cell-cell co-occurrence between two cell types, as represented by the red and blue heatmap colors (Figure 6, F). The degree of overlap between two cell populations in x,y space, represented by the size of heatmap circles, was quantified using the Earth Mover's Distance (Orlova et al., 2016) that is a measure of the distance between two probability distributions. Representative examples for cell-cell co-occurrence and avoidance are shown on (Figure 6, G). Performance of this approach to assess cellular organization is comparable with the CLQ method that we proposed here (Supplementary Fig 4,A). Among the 289 pairs of cell types, 283 pairs from both approaches yielded the same conclusion (Cohen's kappa = 0.953), indicating almost perfect agreement between the two methods. For those 6 inconsistent pairs, no complete opposite conclusion between two methods were observed, i.e. the conclusion of "non-significant" was observed in one of the methods (Supplementary Table 2).

# 3 Discussion

Complex multicellular organisms cells are arranged into tissues that are in turn organized into organs, and so on. At each level of organization, structure is closely related to function which is in turn is governed by different gene expression programs that potentially depend on the cellular microenvironment of an individual cell. This suggests that if we could measure gene and protein expression of individual cells in a given tissue and be able to efficiently incorporate spatial cell arrangement information into these measurements, we would be able to discern the molecular basis of tissue function.

Currently, the fast pace development of spatial transcriptomics and proteomics is bringing us closer to fulfill this scientific program. Nevertheless, the data generated by these methodologies is very rich in information and its successful analysis and interpretation depends on the existence of computational pipelines. Most of the solutions provided by the scientific community are libraries for the use of computational biologists which exposes the need for a graphical user interface solution to enable the direct interaction with the data from other domain experts such as pathologists or biomedical scientists.

SPEX was presented in this work to meet the need of a graphical user interface toolkit for spatially resolved omics. The built in modular nature of SPEX allowed us to construct different data analysis pipelines to study a variety of use case. These pipelines, were informed by the ability of the pathologist to interact with the data through our graphical user interface highlighting the importance of having a tool with these capabilities. We demonstrated SPEX pipelines by identifying tumor-immune microenvironment composition in a set of heterogeneous samples from a cohort of triple negative breast cancer patients. The spatial analytics module was applied to a MERFISH tumor sample where SPEX communities' pathway analysis identified microenvironment domains within the tumor tissue where particular transcriptional programs are more active. Across all these use cases, SPEX showed its potential to discover novel biology from spatially resolved omics data.

SPEX also introduced the novel application of the CLQ methodology to biological tissue samples together with a novel method for spatially resolved proteomics data clustering. We are aware that

there are still many open questions that warrant the development of analysis modules. Our analyses indicate that "ambient RNA" diffusion in the slide is a problem of MERFISH samples and needs to be solved beyond improving on the cell segmentation step. This is an example of many possibilities for future development and, therefore, we hope that the scientific community will embrace this software tool and expand upon its modules.

# 4 Methods

## 4.1 Data Overview

To demonstrate SPEX capabilities, we leverage both newly generated and published datasets. The published datasets include raw MIBI data for a cohort of 41 TNBC patients that was downloaded from `https://www.angelolab.com/mibi-data` (Keren et al., 2018). In the next section, we provide a description of the methodology followed to generate the new datasets presented in this work.

**Control Tonsil Tissue** Antibody staining for imaging mass cytometry was performed on 4$\mu m$-thick, formalin-fixed, paraffin-embedded tissue sections mounted on SuperFrost PlusTM glass slides. Antibodies to cytokeratin (cat. no. ab80826), CD103 (cat. no. ab221210) and PD-L1 (cat. no. ab226766) were purchased as purified, carrier-free formulations from AbCam and, using the Maxpar X8 Antibody Labeling Kit (Fluidigm Inc., South San Francisco, CA), conjugated to 176-Yb (cat. no. 201167A), 153-Eu (cat. no. 201153A) and 150-Nd (cat. no. 201150A), respectively, per manufacturer's protocol. Remaining antibodies were purchased as conjugated from Fluidigm. To stain, tissue was first deparaffinized and rehydrated. Next, pH 9.0 antigen retrieval was used followed by incubation in Superblock at room temperature for 30 minutes. Samples were incubated with a staining panel comprised of 16 antibodies (see Supplementary table1) overnight at 4C in a humidity chamber. Samples were washed four times in 0.1% TritonX/PBS for four minutes each followed by two 4 minute washes in PBS at room temperature. Samples were then incubated with Ir-Intercalator dye for 30 min at room temperature, washed in water for 5 min and air-dried. Regions of interest (ROIs) measuring 1800 x 2000 um each were selected from the slide and acquired at 200 Hz. Data were exported as MCD files and visualized.

**MERFISH Lung Tumor Sample** Samples from patients with human lung cancer were snap frozen and preserved in optimal cutting temperature (OCT) compound and cut into 10 $\mu m$ thick on a cryostat at -20C and placed on MERSCOPE Slide (Vizgen 20400001). The tissue slices were fixed with 4% paraformaldehyde in 1x PBS for 15 minutes, washed three times with 5 mL 1XPBS and incubated with 70% ethanol at 4 C overnight for tissue permeabilization. Samples were then stained for cell boundary using Vizgen's Cell Boundary Kit (10400009), and later hybridized with a custom designed MERSCOPE Gene Panel Mix consisting of 484 genes that assess the different cell types (tumor, stroma, immune cells, etc) and key cell signaling and activity marker (Vizgen 20300008) at 37C incubator for 36-48 hours. Following incubation, the tissues were washed with 5 mL Formamide Wash Buffer at 47C for 30 minutes, twice and embedded into a hydrogel using the Gel Embedding Premix (Vizgen 20300004), ammonium persulfate (Sigma, 09913-100G) and TEMED (N,N,N',N'-tetramethylethylenediamine) (Sigma, T7024-25ML) from the MERSCOPE Sample Prep Kit (10400012). After the gel mix solution solidified, the samples were cleared with Clearing Solution consisting of 50uL of Protease K (NEB, P8107S) and 5mL of Clearing Premix (Vizgen 20300003) at 37C overnight. After removing clearing solution,

the sample was stained with DAPI and Poly T Reagent (Vizgen 20300021) for 15 minutes at room temperature, washed for 10 minutes with 5ml of Formamide Wash Buffer, and then imaged on the MERSCOPE system (Vizgen 10000001). A fully detailed, step-by-step instruction on the MERFISH sample prep the full protocol is available at https://vizgen.com/resources/fresh-and-fixed-frozen-tissue-sample-preparation/. Full Instrumentation protocol is available at https://vizgen.com/resources/merscope-instrument/.

## 4.2 SPEX modules

The SPEX graphical user interface provides a subset of essential components which will facilitate end to end analysis of FOV (Field of View) images. Additional algorithms with source code are made available in the code repository. The source code provided allows researchers to scale analytical approaches to WSI (Whole Slide Images) and integrate with the other workflows. Code will be made available upon publication.

### 4.2.1 Image processing

Image processing is executed in a 5 step sequence with the ultimate goal of generating a cell by expression matrix. These steps include- image loading, image preprocessing, single-cell segmentation, post-processing and feature extraction. Each step contains a selection of modules which can be linked together to address the particularities of a given image set.

**Image loading** The image loading step contains a single module which supports OME-TIFF. This multi-dimensional open format provides a well-structured metadata header that can accommodate a wide range of image and acquisition system information. A number of open source conversion pipelines are available in the event proprietary microscopy formats need to be converted to OMETIFF. Within SPEX, the TIFFFILE Python library is used to transform the OME-TIFF image to an in-memory multichannel Numpy array. OMETIFF channel and resolution metadata is parsed by the AICSimageio python library and is used for downstream processing steps (https://github.com/AllenCellModeling/aicsimageio).

**Preprocessing** The preprocessing step includes optional modules to denoise images and/or enhance pixel information to facilitate single-cell segmentation. These modules include- global background correction, median filter denoising, and Non-local means (NLM) denoising. All modules were developed using the Scikit-image python library. To accommodate the potentially large channel dimension of images, channels are processed in parallel with use of the apply_parallel() function in SCIKIT image, a wrapper for DASK *map_blocks()* function. This routine allows application of the denoising functions on chunked arrays which are efficiently distributed across computing cores.

In the global background correction module, background signal, as captured in one channel, can be subtracted from all other channels. For example, signal in an autofluorescent or a detector noise channel can be subtracted from channels that house specific molecular markers. Within the defined background channel, OTSU thresholding (Otsu, 1979) is applied to create a positive pixel binary mask. This pixel mask is then intensity-scaled by a user-defined correction factor. This correction mask is then subtracted from the other channels.

Median denoising is a common image processing technique which can preserve edge information while suppressing noise. For any channel, the filter replaces a pixel's value with the median of

its local neighborhood. The size of the local neighborhood is defined by the filter kernel size, an argument of the function.

Non-local means denoising provides a slightly more sophisticated method where a pixel's value is replaced by a mean which is sampled from other regions of the image. These sampled regions are only utilized if their mean is similar to the target region. This technique can preserve local texture information.

**Single-cell segmentation** SPEX includes several algorithms for single-cell segmentation. This includes traditional watershed cell segmentation in addition to a collection of pretrained deep learning models which include Stardist (Schmidt et al., 2018), Cellpose (Stringer et al., 2021) and Mesmer (Greenwald et al., 2022). These algorithms will operate on the channel or combination of channels containing cell nucleus information. Cellpose and Mesmer will also accept cytoplasmic or membrane marker channels as input. In instances where the input data differs in resolution to the training dataset for the respective models, the user can up-sample or downsample the image pixels. All segmentation modules will return an instance segmentation label image where each cell is assigned a unique integer value.

**Post-processing** The optional postprocessing modules aim to modify the single-cell segmentation labels. Tissue-based imaging is often prone to include technical artifacts. This could be tissue tears, folds or debris. These artifacts may generate false positive cell segmentation labels. Therefore, we include rule-based functions to exclude segmented labels based on morphological features.

In some rare instances, the pretrained deep learning segmentation may include false negative regions. This is likely a result of the model not having representation of this cell type in the training data. In these cases, one can utilize the optional *rescue_cells()* algorithm . First, prototypical cell size and intensity are calculated adaptively from the currently segmented cell objects. These parameters are then used as input for classical watershed segmentation of the image. If a label from the watershed segmentation does not overlap with that of the deep learning approach, then it is merged into the final instance segmentation label image.

The user then has an option to expand the boundaries of the cell. This may be desired since some of the cell segmentation modules utilize only nuclear information. To ensure cytoplasmic compartments are included in feature extraction, boundaries should be dilated.

**Feature extraction** SPEX utilizes the scikit image regionprops function to extract single cell features. Currently, these features include mean intensity for each channel, extracted for each cell. The output will be a cell by expression matrix which serves as input for downstream phenotyping.

### 4.2.2 Proteomics Data Clustering

A cell by expression matrix obtained from the set of image processing modules can be clustered in a variety of ways into the groups of cells that share the same properties. Currently we provide two options to do clustering: bulk (multiple samples clustered together) and per sample with the further cluster alignment between samples.

**Bulk Clustering** The cell by expression matrix was transformed with arcsine cofactor 5 and then standardized prior to clustering with Phenograph using the recommended default user input parameter settings (kNN=30)(Levine et al., 2015).

**Bulk Clustering** This clustering approach requires a representative training set that can be built from randomly subsampling from the collection of samples of a particular pool. Alternatively, one or a few reference samples could serve as a training set. Each sample from the pool is then used as an independent test set and processed separately. We first do per sample transformation (arcsine cofactor 5) and standardization. The training set is pre-processed separately from each of the test sets. The training set is then clustered with Phenograph and the obtained cluster ids are then used to build the supervised embedding with UMAP (McInnes et al., 2018)( `https://umap-learn.readthedocs.io/en/latest/supervised.html`) that separate labeled classes best. Each test sample is then embedded together with the training set using the metric learning with UMAP. Each test set is then clustered with the DBSCAN directly in that embedded space and QFMatch (Orlova et al., 2018) is used to align cluster ids between the test set and the training set. Phenotypic labels are assigned based on exploring the marker expression pattern and spatial location for each of the automatically identified clusters.

#### 4.2.3 Spatial Analytics Module

**Cell-cell co-occurrence at the cell population level** To systematically assess the non-random co-occurrence/avoidance of identified cell types at the cell population level, we used a permutation test to compare the number of interactions between all of the cell types in a given image (or in a given user defined region) to that of synthetic matched controls generated under the null hypothesis containing randomized cell phenotypes. The neighborhood radius (e.g. 20 $\mu m$) was chosen empirically to restrict cell-cell co-occurrence assessment to only those cells that are close enough to potentially establish physical contact between their membranes. This approach, similar to the one described in the histoCAT paper (Schapiro et al., 2017), allows to determine the significance of cell-cell co-occurrence between two cell types. The degree of overlap between two cell populations in x,y space, was quantified using the Earth Mover's Distance (Orlova et al., 2016) that is a measure of the distance between two probability distributions.

**Colocation Quotient: cell-cell co-occurrence at enhanced resolution** The Colocation Quotient (CLQ), defined as a ratio of ratios, is a widely used geographic index (Leslie and Kronenfeld, 2011). In the context of tissue images, it is used to measure the local density to the global density of a specific cell type. Specifically, the local density is calculated as the proportion of cell type B in the neighborhood constructed by a certain radius, centered around cell type A. The global density is the proportion of cell type B in the entire tissue image. In general, if the density of cell type B within the neighborhood of A is more than the global density of B, the CLQ will be $> 1$. If the neighborhood of cell type A contains many other cell types other than B, the CLQ will be $< 1$. The value 1 means there is no spatial relationships between the two cell types.

The CLQ can be measured globally for a particular cell type A, $\mathrm{CLQ}_{A \to B}$, or locally for each cell in the slide classified as that particular cell type, this we denote as $\mathrm{LCLQ}_{A_i \to B}$. The CLQ is

calculated through the following expressions:

$$\mathrm{LCLQ}_{A_i \to B} = \frac{N_{A_i \to B}}{N_b\,(N-1)} \tag{1}$$

$$N_{A_i \to B} = \sum \frac{w_{ij}\delta_{ij}}{\sum w_{ij}} \tag{2}$$

$$\mathrm{CLQ}_{A \to B} = \frac{\sum_{i=1}^{N_A} \mathrm{LCLQ}_{A_i \to B}}{N_A} \tag{3}$$

Where $N$ is the total number of cells in the image, $\delta_{ij}$ is the Kroenecker delta indicating whether cell $j$ is a type B cell; $w_{ij}$ is $1/N$ for non-weighted version, and Gaussian distance decay kernel for weighted version which is common in geostatistics application (Li et al., 2011). The global CLQ will then be the average of the local CLQ.

**CLQ for tumor immunophenotyping** We further developed a two-step approach based on the proposed CLQ scores to quantify the degree of mixing between tumor and immune cells. Firstly, we set a hard-cutoff threshold to determine the immune desert class, any sample that has less than 200 immune cells will be regarded as a desert sample; in the second step, using the rest of the samples, we further implemented a two-component Gaussian-mixture model that utilized the $CLQ_{immune \to tumor}$ and $CLQ_{tumor \to immune}$ to distinguish between immune excluded and inflamed categories. Intuitively, the inflamed category will have a larger mean of CLQ scores than the excluded category.

**Spatially informed transcriptomics using CLQ scores** For each cell type we estimated a CLQ score that describes the spatial relationship between cells (or globally as clusters), classifying them as "avoidant" or "attractive" spatial relationships. The cell-level CLQ classifications can be further investigated to identify differences in gene expression associated with for example "avoidant" behavior in clusters of interest. Gene expression differences were estimated using the Seurat R package (version 4.0.4) (Hao et al., 2021) using default parameters. Hypothesis-driven testing of cell clusters and cell-cell spatial relationships for differences in gene expression can be performed flexibly using SPEX. Additional, parameter specifications to the *Seurat::FindMarkers()* function are also possible (for example to allow for using different statistical tests).

**Spatially informed transcriptomics using SPEX clusters** For each cell in the dataset, neighborhood compositions were determined by collecting the counts of cells belonging to the different clusters included within a circle of radius epsilon centered on each cell (a parameter that can be adjusted by the user to assess the locality of expected cell interactions). SPEX communities then were identified via clustering of the matrix of the neighborhood compositions using Phenograph.

MERFISH single cell data was processed using the Seurat pipeline(version 4.0.4) and then fed into SPEX for spatial analyses. For each cluster identified, CLQ was calculated and neighborhood compositions were determined for each cell. Spex communities were identified via clustering neighborhood compositions using hdbscan. These communities were then analyzed using CellChat (Jin et al., 2021) (version 1.1.0) to infer cell-cell interactions and pathways specific for said communities. SPEX communities were also inspected for pathway activities from a collection of well understood transcriptional pathways using progeny (version 1.12.0) (Schubert et al., 2018).

## References and Notes


Allen, C. D., Okada, T., and Cyster, J. G. (2007). Germinal-center organization and cellular dynamics. *Immunity*, 27(2):190–202.

Bannard, O. and Cyster, J. G. (2017). Germinal centers: programmed for affinity maturation and antibody diversification. *Current opinion in immunology*, 45:21–30.

Chen, D. S. and Mellman, I. (2017). Elements of cancer immunity and the cancer–immune set point. *Nature*, 541(7637):321–330.

Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., and Zhuang, X. (2015). Spatially resolved, highly multiplexed rna profiling in single cells. *Science*, 348(6233):aaa6090.

Dries, R., Chen, J., Del Rossi, N., Khan, M. M., Sistig, A., and Yuan, G.-C. (2021). Advances in spatial transcriptomic data analysis. *Genome Research*, 31(10):1706–1718.

Galon, J. and Bruni, D. (2019). Approaches to treat immune hot, altered and cold tumours with combination immunotherapies. *Nature reviews Drug discovery*, 18(3):197–218.

Giesen, C., Wang, H. A., Schapiro, D., Zivanovic, N., Jacobs, A., Hattendorf, B., Schüffler, P. J., Grolimund, D., Buhmann, J. M., Brandt, S., et al. (2014). Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature methods*, 11(4):417–422.

Goltsev, Y., Samusik, N., Kennedy-Darling, J., Bhate, S., Hale, M., Vazquez, G., Black, S., and Nolan, G. P. (2018). Deep profiling of mouse splenic architecture with codex multiplexed imaging. *Cell*, 174(4):968–981.

Greenwald, N. F., Miller, G., Moen, E., Kong, A., Kagel, A., Dougherty, T., Fullaway, C. C., McIntosh, B. J., Leow, K. X., Schwartz, M. S., et al. (2022). Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nature biotechnology*, 40(4):555–565.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck III, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.

Jin, S., Guerrero-Juarez, C. F., Zhang, L., Chang, I., Ramos, R., Kuan, C.-H., Myung, P., Plikus, M. V., and Nie, Q. (2021). Inference and analysis of cell-cell communication using cellchat. *Nature communications*, 12(1):1–20.

Johnson, D. B., Bordeaux, J., Kim, J. Y., Vaupel, C., Rimm, D. L., Ho, T. H., Joseph, R. W., Daud, A. I., Conry, R. M., Gaughan, E. M., et al. (2018). Quantitative spatial profiling of pd-1/pd-l1 interaction and hla-dr/ido-1 predicts improved outcomes of anti–pd-1 therapies in metastatic melanoma. *Clinical Cancer Research*, 24(21):5250–5260.

Keren, L., Bosse, M., Marquez, D., Angoshtari, R., Jain, S., Varma, S., Yang, S.-R., Kurian, A., Van Valen, D., West, R., et al. (2018). A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging. *Cell*, 174(6):1373–1387.

Leslie, T. F. and Kronenfeld, B. J. (2011). The colocation quotient: A new measure of spatial association between categorical subsets of points. : . *Geographical Analysis*, 43(3):306–326.

Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L., El-ad, D. A., Tadmor, M. D., Litvin, O., Fienberg, H. G., Jager, A., Zunder, E. R., et al. (2015). Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1):184–197.

Lewis, S. M., Asselin-Labat, M.-L., Nguyen, Q., Berthelet, J., Tan, X., Wimmer, V. C., Merino, D., Rogers, K. L., and Naik, S. H. (2021). Spatial omics and multiplexed imaging to explore cancer biology. *Nature methods*, 18(9):997–1012.

Li, B., Gould, J., Yang, Y., Sarkizova, S., et al. (2022). Pegasus documentation.

Li, X., Sun, X., Jin, L., and Xue, F. (2011). Worldwide spatial genetic structure of angiotensin-converting enzyme gene: a new evolutionary ecological evidence for the thrifty genotype hypothesis. *European journal of human genetics*, 19(9):1002–1008.

Liu, H., Liu, L., Liu, K., Bizargity, P., Hancock, W. W., and Visner, G. A. (2009). Reduced cytotoxic function of effector cd8+ t cells is responsible for indoleamine 2, 3-dioxygenase-dependent immune suppression. *The Journal of Immunology*, 183(2):1022–1031.

Martinelli, A. L. and Rapsomaniki, M. A. (2022). Athena: Analysis of tumor heterogeneity from spatial omics measurements. *Bioinformatics*.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Orlova, D. Y., Meehan, S., Parks, D., Moore, W. A., Meehan, C., Zhao, Q., Ghosn, E. E., Herzenberg, L. A., and Walther, G. (2018). Qfmatch: multidimensional flow and mass cytometry samples alignment. *Scientific reports*, 8(1):1–14.

Orlova, D. Y., Zimmerman, N., Meehan, S., Meehan, C., Waters, J., Ghosn, E. E., Filatenkov, A., Kolyagin, G. A., Gernez, Y., Tsuda, S., et al. (2016). Earth mover's distance (emd): a true metric for comparing biomarker expression levels in cell populations. *PloS one*, 11(3):e0151859.

Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66.

Palla, G., Fischer, D. S., Regev, A., and Theis, F. J. (2022a). Spatial components of molecular tissue biology. *Nature Biotechnology*, pages 1–11.

Palla, G., Spitzer, H., Klein, M., Fischer, D., Schaar, A. C., Kuemmerle, L. B., Rybakov, S., Ibarra, I. L., Holmberg, O., Virshup, I., et al. (2022b). Squidpy: a scalable framework for spatial omics analysis. *Nature methods*, 19(2):171–178.

Schapiro, D., Jackson, H. W., Raghuraman, S., Fischer, J. R., Zanotelli, V. R., Schulz, D., Giesen, C., Catena, R., Varga, Z., and Bodenmiller, B. (2017). histocat: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nature methods*, 14(9):873–876.

Schmidt, U., Weigert, M., Broaddus, C., and Myers, G. (2018). Cell detection with star-convex polygons. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 265–273. Springer.

Schubert, M., Klinger, B., Klünemann, M., Sieber, A., Uhlitz, F., Sauer, S., Garnett, M. J., Blüthgen, N., and Saez-Rodriguez, J. (2018). Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nature communications*, 9(1):1–11.

Stringer, C., Wang, T., Michaelos, M., and Pachitariu, M. (2021). Cellpose: a generalist algorithm for cellular segmentation. *Nature methods*, 18(1):100–106.

Tsakiroglou, A. M., Fergie, M., Oguejiofor, K., Linton, K., Thomson, D., Stern, P. L., Astley, S., Byers, R., and West, C. M. (2020). Spatial proximity between t and pd-l1 expressing cells as a prognostic biomarker for oropharyngeal squamous cell carcinoma. *British journal of cancer*, 122(4):539–544.

Willard-Mack, C. L. (2006). Normal structure, function, and histology of lymph nodes. *Toxicologic pathology*, 34(5):409–424.

# 5 Figures

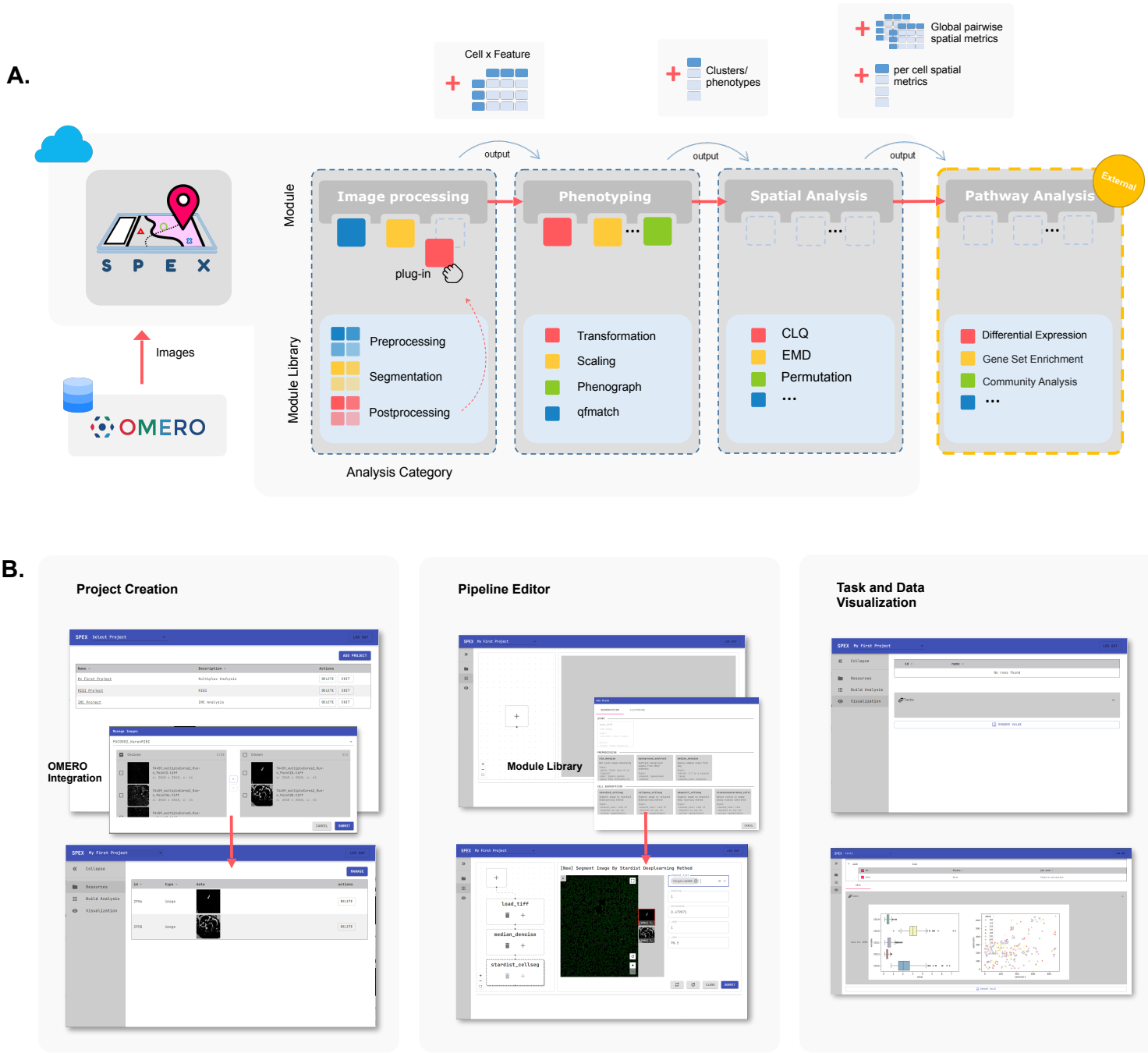
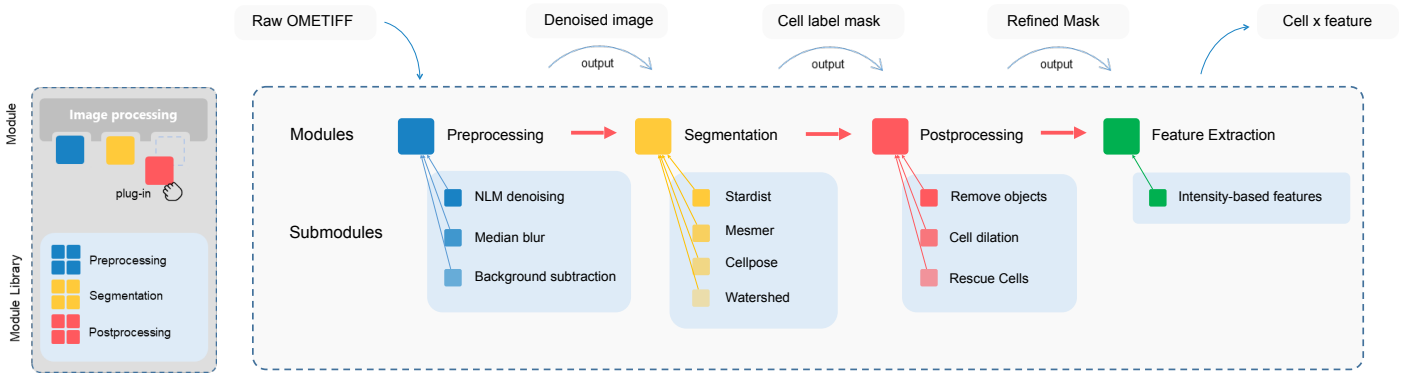

A.
Cell x Feature
Clusters/ phenotypes
Global pairwise spatial metrics
per cell spatial metrics
S P E X
Images
OMERO
Module
Module Library
output
output
output
Image processing
plug-in
Preprocessing
Segmentation
Postprocessing
Phenotyping
Transformation
Scaling
Phenograph
qfmatch
Spatial Analysis
CLQ
EMD
Permutation
...
Pathway Analysis
External
Differential Expression
Gene Set Enrichment
Community Analysis
...
Analysis Category
B.
Project Creation
OMERO Integration
Pipeline Editor
Module Library
Task and Data Visualization

**Figure 1: SPEX graphical user interface and built-in modularity enables user defined pipelines.** A: Overview of SPEX workflow showcasing the different modules available within each analysis step. **B:** Snapshots of the SPEX graphical user interface during several steps of the analysis pipeline.

**A.**


Raw OMETIFF
Denoised image
output
Cell label mask
output
Refined Mask
output
Cell x feature
Module
Image processing
plug-in
Module Library
Preprocessing
Segmentation
Postprocessing
Modules
Preprocessing
Segmentation
Postprocessing
Feature Extraction
Submodules
NLM denoising
Median blur
Background subtraction
Stardist
Mesmer
Cellpose
Watershed
Remove objects
Cell dilation
Rescue Cells
Intensity-based features

**B.**

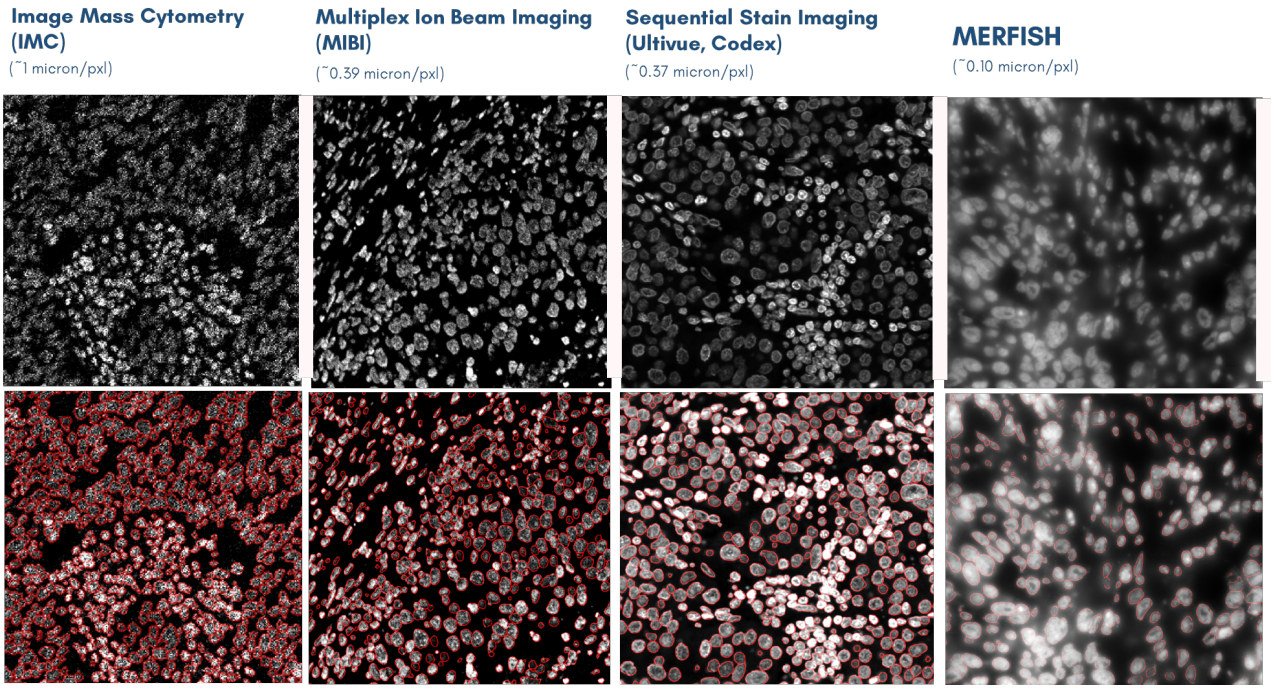
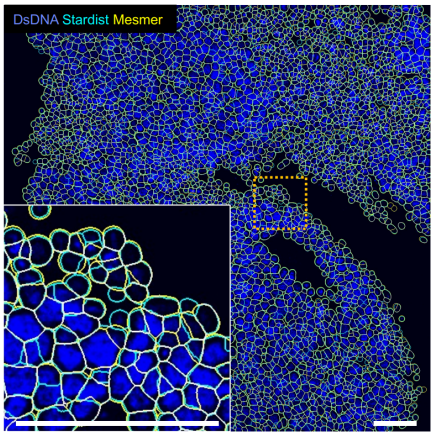


Image Mass Cytometry (IMC)
(~1 micron/pxl)
Multiplex Ion Beam Imaging (MIBI)
(~0.39 micron/pxl)
Sequential Stain Imaging (Ultivue, Codex)
(~0.37 micron/pxl)
MERFISH
(~0.10 micron/pxl)

**C.**


DsDNA Stardist Mesmer

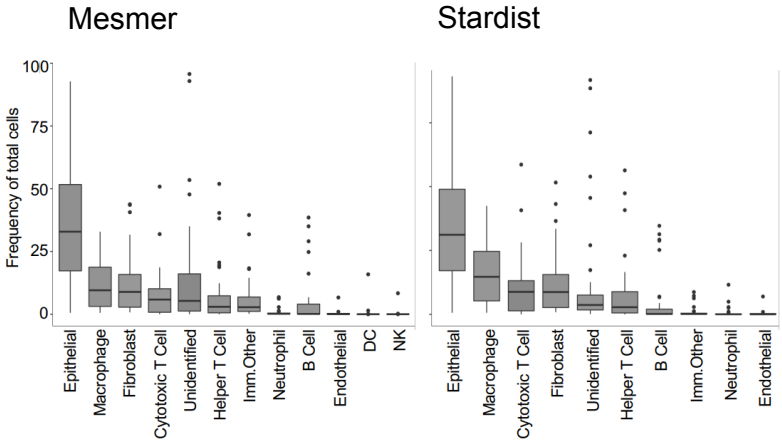

Mesmer
Stardist
Frequency of total cells
100
75
50
25
0
Epithelial
Macrophage
Fibroblast
Cytotoxic T Cell
Unidentified
Helper T Cell
Imm.Other
Neutrophil
B Cell
Endothelial
DC
NK

**Figure 2: SPEX Image processing module. A:** SPEX image processing workflow details. Images are load, preprocessed, segmented and post processed in order to enable feature extraction. **B:**Examples of imaging modalities segmented using SPEX showing a variety of resolutions and noise. **C:** Comparison of pre-trained deep learning segmentation models. Left, Color overlay showing DsDNA-positive cell nuclei (blue) in a MIBI-TOF image with overlaid segmentation outlines from Stardist (cyan) and Mesmer (yellow) segmentation workflows. Boxplots showing cell type abundance across the entire 41-patient MIBI-TOF breast cancer following single cell annotation in SPEX using either the Stardist of Mesmer segmentation results.

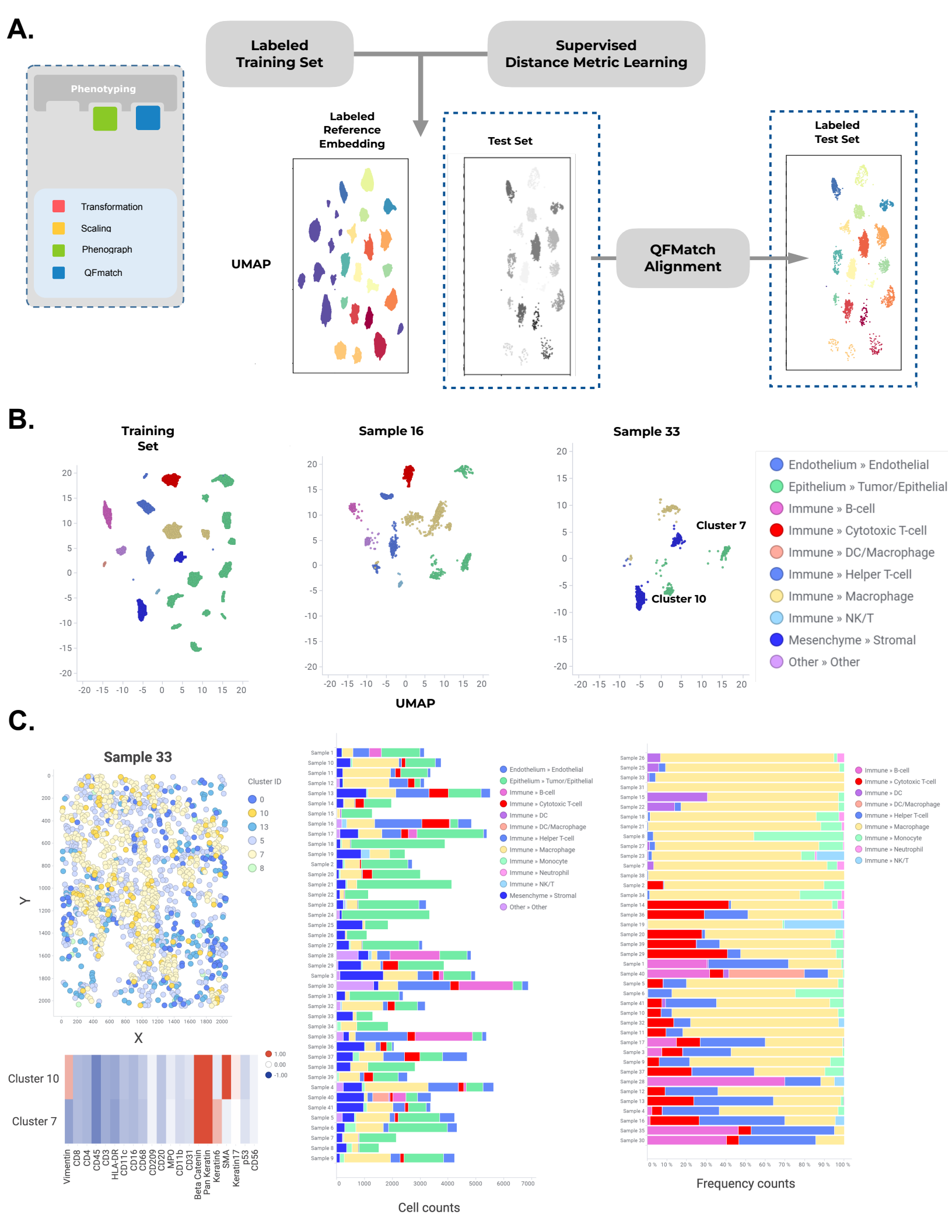

A.
Phenotyping
Transformation
Scaling
Phenograph
QFmatch
Labeled Training Set
Supervised Distance Metric Learning
Labeled Reference Embedding
UMAP
Test Set
QFMatch Alignment
Labeled Test Set
B.
Training Set
Sample 16
Sample 33
Cluster 7
Cluster 10
UMAP
Endothelium » Endothelial
Epithelium » Tumor/Epithelial
Immune » B-cell
Immune » Cytotoxic T-cell
Immune » DC/Macrophage
Immune » Helper T-cell
Immune » Macrophage
Immune » NK/T
Mesenchyme » Stromal
Other » Other
C.
Sample 33
Cluster ID
0
10
13
5
7
8
X
Y
Cluster 10
Cluster 7
Vimentin
CD8
CD4
CD45
CD3
HLA-DR
CD11c
CD16
CD68
CD209
CD20
MPO
CD11b
CD31
Beta Catenin
Pan Keratin
Keratin6
SMA
Keratin17
p53
CD56
Cell counts
Frequency counts
Immune » B-cell
Immune » Cytotoxic T-cell
Immune » DC
Immune » DC/Macrophage
Immune » Helper T-cell
Immune » Macrophage
Immune » Monocyte
Immune » Neutrophil
Immune » NK/T

**Figure 3: SPEX Clustering Module A:** SPEX Clustering Module integration algorithm. A reference embedding is generated with the labeled training set using UMAP metric learning. The test sample is then matched into this embedding using QFMatch. **B:** SPEX integration using an annotated reference allows for cell type identification in two unlabelled queried datasets. In sample 33, tumor cells are classified into two different clusters, cluster 7 and 10. **C:** Top, The cluster ids determined by the integration algorithm projected back into the spatial coordinates of sample 33. Bottom, Protein expression for the two tumor populations in said sample. A stromal signature is observed in cluster 10 concordant with their proximity to stroma. **D:** Left, Total counts of each cell ypes detected. Right, Relative abundance of the immune cell compartment.

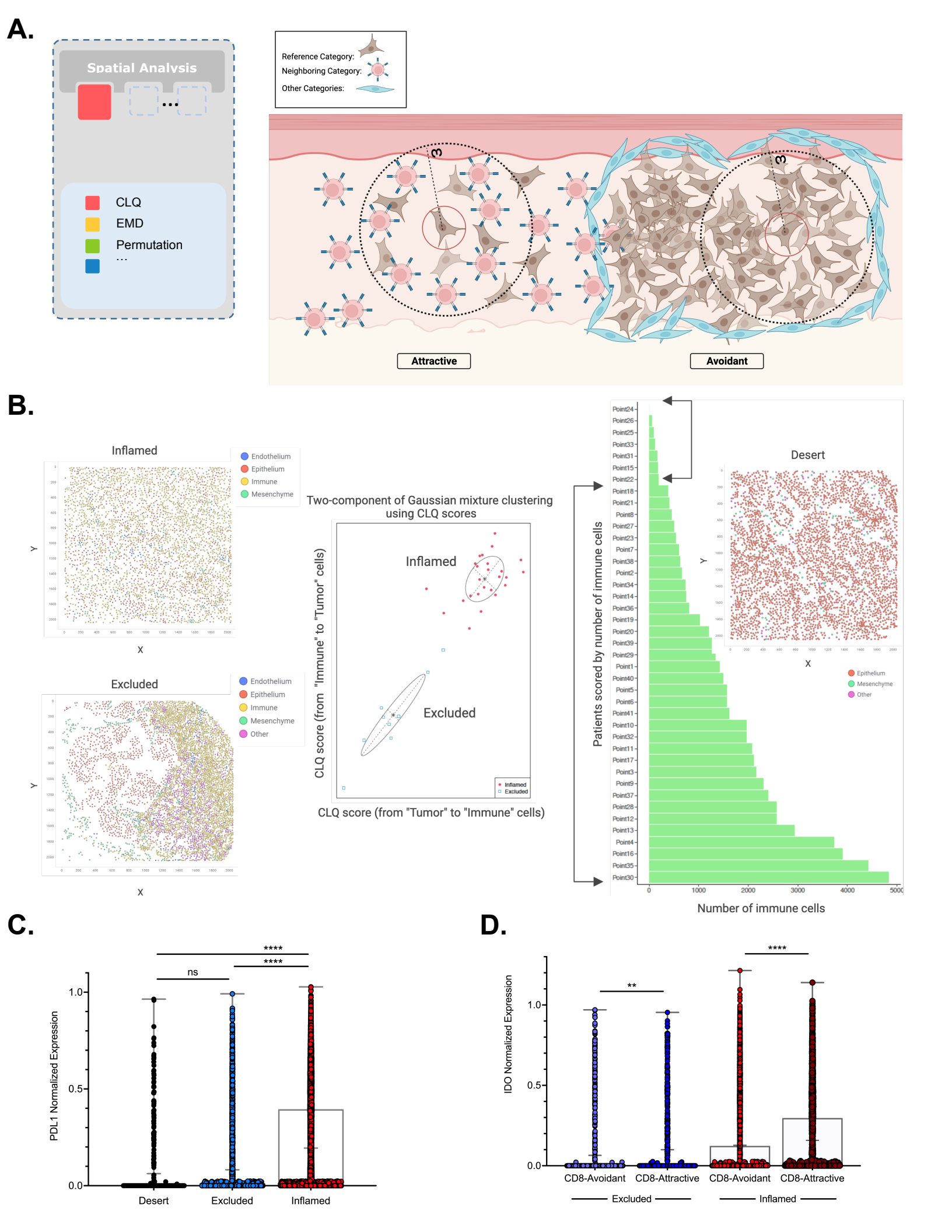

A.
Spatial Analysis
...
CLQ
EMD
Permutation
...
Reference Category:
Neighboring Category:
Other Categories:
3
Attractive
Avoidant
B.
Inflamed
Endothelium
Epithelium
Immune
Mesenchyme
Excluded
Other
X
Y
Two-component of Gaussian mixture clustering using CLQ scores
CLQ score (from "Immune" to "Tumor" cells)
CLQ score (from "Tumor" to "Immune" cells)
Patients scored by number of immune cells
Number of immune cells
Desert
C.
PDL1 Normalized Expression
ns
****
Desert
Excluded
Inflamed
D.
IDO Normalized Expression
**
CD8-Avoidant
CD8-Attractive
Excluded
Inflamed

**Figure 4: CLQ detects avoidance or interaction relationships between cell types in tumor immunophenotyping. A:** Colocation Quotient (CLQ) Method: If the density of cell type B within the locality of A is more than the global density of B, the CLQ will be $> 1$. If the locality of cell type A contains many other cell types other than B, the CLQ will be $< 1$.The value 1 means there is no spatial relationships between the two cell types. The CLQ value can be calculated at each type A cell given a radius to define its locality. This is referred to as local CLQ (LCLQ). According to the local CLQ value, each type A cell can be categorized as "attractive" or "avoidant". Furthermore, the average of all local CLQ will render the global CLQ score to represent the overall spatial relationship from cell type A to cell type B. **B:** CLQ for tumor immunophenotyping. With thresholding and two-component Gaussian mixture modeling, we categorized each sample into corresponding immunophenotyping class, representative samples are shown per class. Overall, there are 33 samples (out of 41) that agree between keren's call and our prediction.**C:**Single cell, normalized expression of PDL1 is examined in macrophages from tumors classified as Desert, Excluded, and Inflamed by CLQ analysis; asterisks denote significant difference in expression, $****P < 0.0001$, ns = not significant, multiple comparison corrected Kruskal Wallis test.**D:** Single cell, normalized expression of IDO1 is examined in macrophages that are spatially enriched with CD8 Cytotoxic T Cells (CD8-Attractive) or spatially distinct (CD8-avoidant) in Excluded and Inflamed tumors, asterisks denote significant difference in expression, $**P < 0.01, ****P < 0.0001$, multiple comparison corrected Kruskal Wallis test.

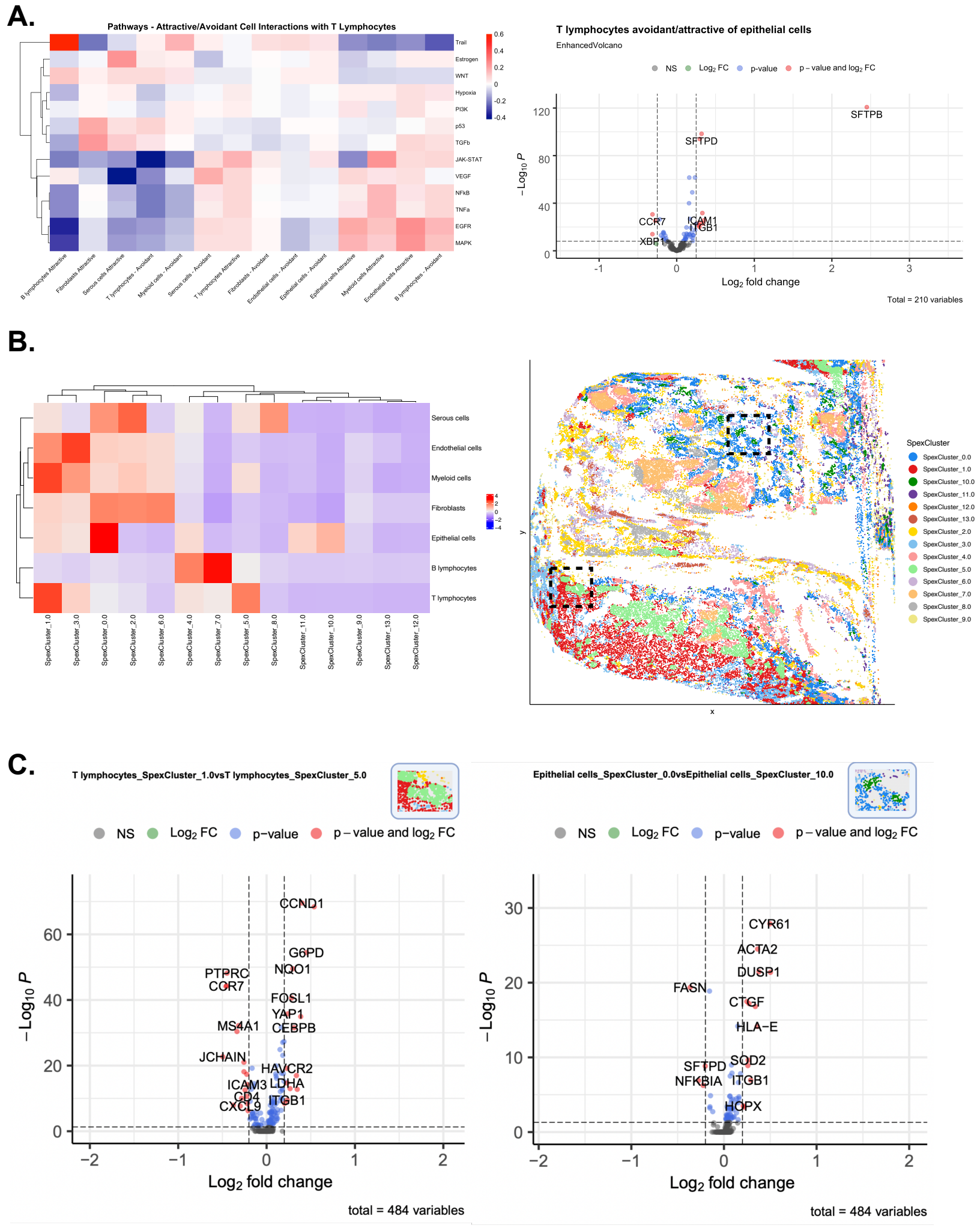
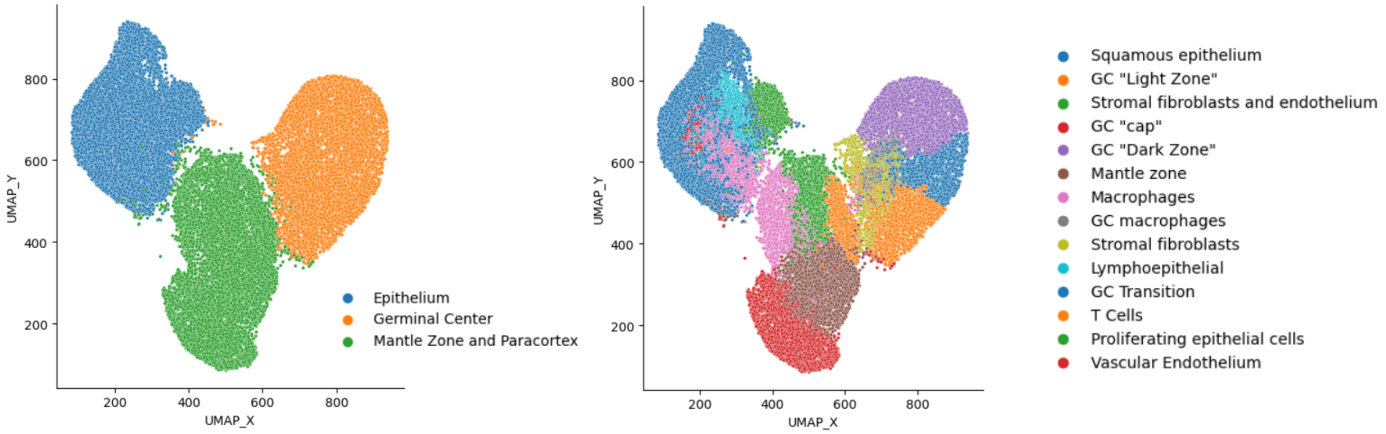
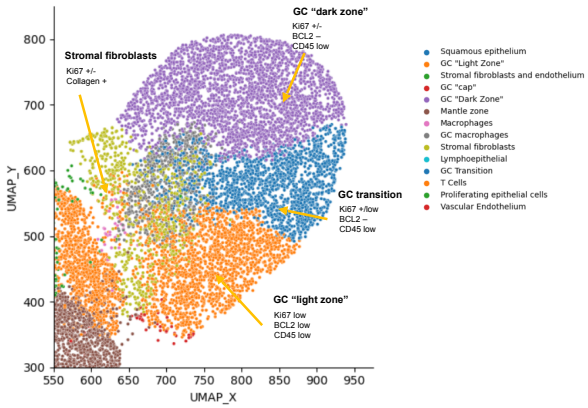

A.
Pathways - Attractive/Avoidant Cell Interactions with T Lymphocytes
Trail
Estrogen
WNT
Hypoxia
PI3K
p53
TGFb
JAK-STAT
VEGF
NFkB
TNFa
EGFR
MAPK
B lymphocytes Attractive
Fibroblasts Attractive
Serous cells Attractive
T lymphocytes - Avoidant
Myeloid cells - Avoidant
Serous cells - Avoidant
T lymphocytes Attractive
Fibroblasts - Avoidant
Endothelial cells - Avoidant
Epithelial cells - Avoidant
Epithelial cells Attractive
Myeloid cells Attractive
Endothelial cells Attractive
B lymphocytes - Avoidant
T lymphocytes avoidant/attractive of epithelial cells
EnhancedVolcano
NS
Log2 FC
p-value
p – value and log2 FC
SFTPB
SFTPD
CCR7
ICAM1
ITGB1
XBP1
−Log10 P
Log2 fold change
Total = 210 variables
B.
Serous cells
Endothelial cells
Myeloid cells
Fibroblasts
Epithelial cells
B lymphocytes
T lymphocytes
SpexCluster_1.0
SpexCluster_3.0
SpexCluster_0.0
SpexCluster_2.0
SpexCluster_6.0
SpexCluster_4.0
SpexCluster_7.0
SpexCluster_5.0
SpexCluster_8.0
SpexCluster_11.0
SpexCluster_10.0
SpexCluster_9.0
SpexCluster_13.0
SpexCluster_12.0
SpexCluster
SpexCluster_0.0
SpexCluster_1.0
SpexCluster_10.0
SpexCluster_11.0
SpexCluster_12.0
SpexCluster_13.0
SpexCluster_2.0
SpexCluster_3.0
SpexCluster_4.0
SpexCluster_5.0
SpexCluster_6.0
SpexCluster_7.0
SpexCluster_8.0
SpexCluster_9.0
y
x
C.
T lymphocytes_SpexCluster_1.0vsT lymphocytes_SpexCluster_5.0
NS
Log2 FC
p−value
p – value and log2 FC
CCND1
G6PD
NQO1
PTPRC
CCR7
FOSL1
YAP1
MS4A1
CEBPB
JCHAIN
HAVCR2
ICAM3
LDHA
CD4
ITGB1
CXCL9
−Log10 P
Log2 fold change
total = 484 variables
Epithelial cells_SpexCluster_0.0vsEpithelial cells_SpexCluster_10.0
NS
Log2 FC
p−value
p – value and log2 FC
CYR61
ACTA2
DUSP1
FASN
CTGF
HLA-E
SOD2
SFTPD
NFKBIA
ITGB1
HOPX
−Log10 P
Log2 fold change
total = 484 variables

**Figure 5: SPEX pathway analysis module guides biological interpretation through spatially informed differential expression and pathway analysis of a MERFISH lung tumor sample. A.** Left, PROGENy pathway analysis of each cell type in the lung tumor sample stratified by their behavior of either avoidance or attraction towards T cells. Right, Volcano plot showing differentially expressed genes between T cells that are avoidant to tumor cells and those that are attractive. **B**. Left, A heatmap of normalized cell abundances conforming each of the SPEX communities. Right, SPEX communities and their location in the tissue slide. The dashed squares highlight areas with examples the SPEX communities that were analyzed further for contextual differential gene expression. **C.** Volcano plots showing how T cells and exhibit differentially expressed genes in the context of two different SPEX communities. The squares on top correspond to the previously defined areas of the tumor slide.

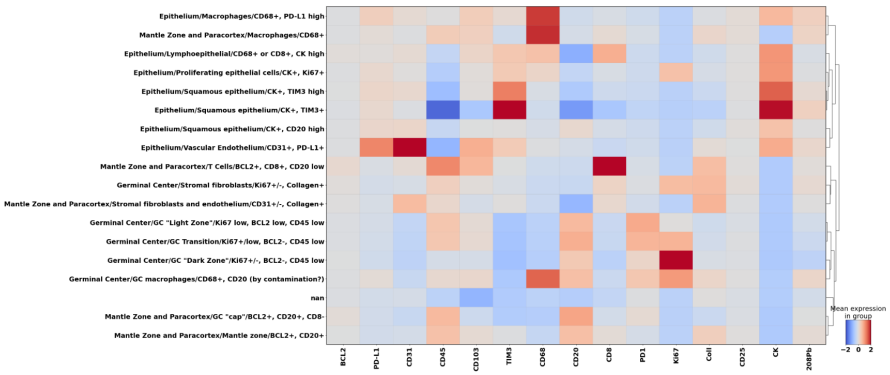
**A.**



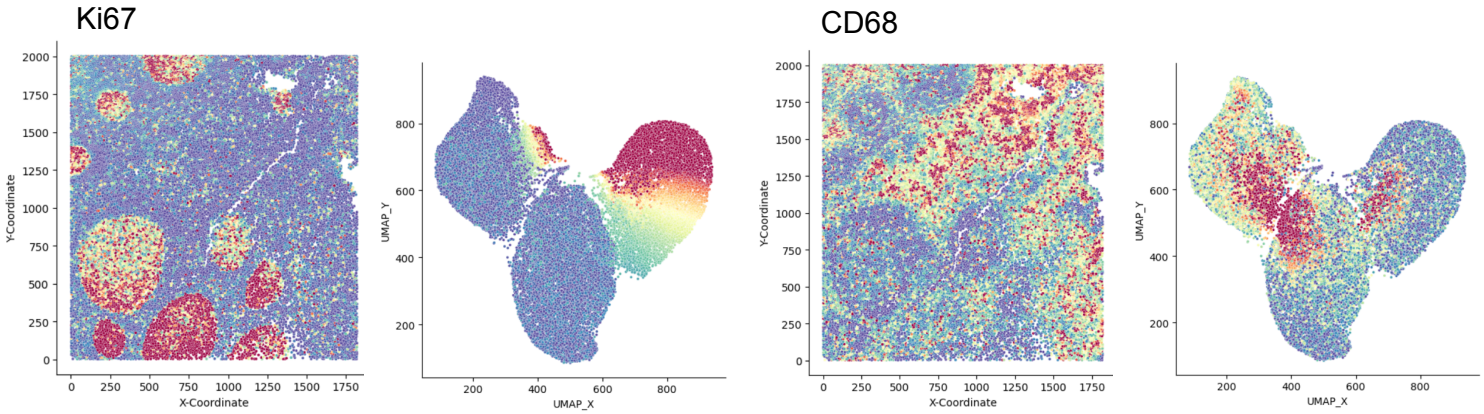

**B.**



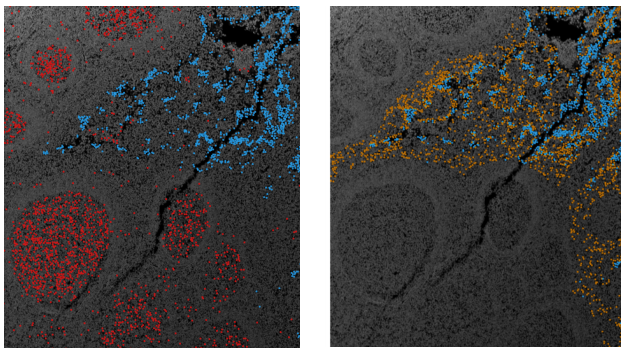

**C.**

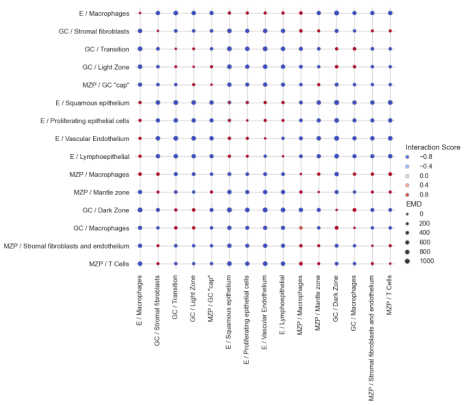



**D.**




**E.**



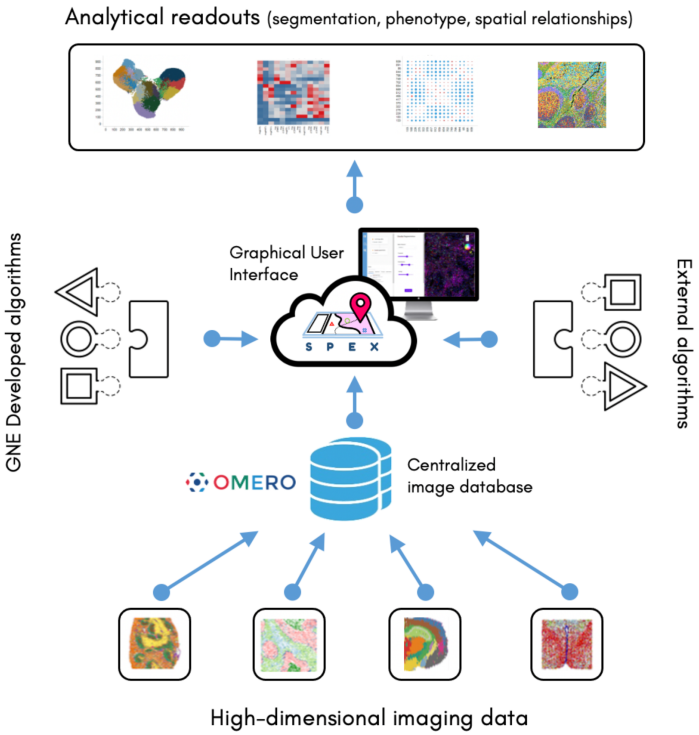




| Zone | Classification | Expression Pattern |
|---|---|---|
| Germinal Center | Stromal fibroblasts | Ki67+/-, Collagen+ |
| Germinal Center | Dark Zone | Ki67+/-, BCL2-, CD45 low |
| Germinal Center | Transition | Ki67+/low, BCL2-, CD45 low |
| Germinal Center | Light Zone | Ki67 low, BCL2 low, CD45 low |
| Germinal Center | Macrophages | CD68+, CD20 (by contamination?) |
| Mantle Zone and Paracortex | Macrophages | CD68+ |
| Mantle Zone and Paracortex | Stromal fibroblasts and endothelium | CD31+/-, Collagen+ |
| Mantle Zone and Paracortex | T Cells | BCL2+, CD8+, CD20 low |
| Mantle Zone and Paracortex | Mantle zone | BCL2+, CD20+ |
| Mantle Zone and Paracortex | GC "cap" | BCL2+, CD20+, CD8- |
| Epithelium | Squamous epithelium | CK+, TIM3+ |
| Epithelium | Squamous epithelium | CK+, TIM3 high |
| Epithelium | Vascular Endothelium | CD31+, PD-L1+ |
| Epithelium | Squamous epithelium | CK+, CD20 high |
| Epithelium | Proliferating epithelial cells | CK+, Ki67+ |
| Epithelium | Lymphoepithelial | CD68+ or CD8+, CK high |
| Epithelium | Macrophages | CD68+, PD-L1 high |

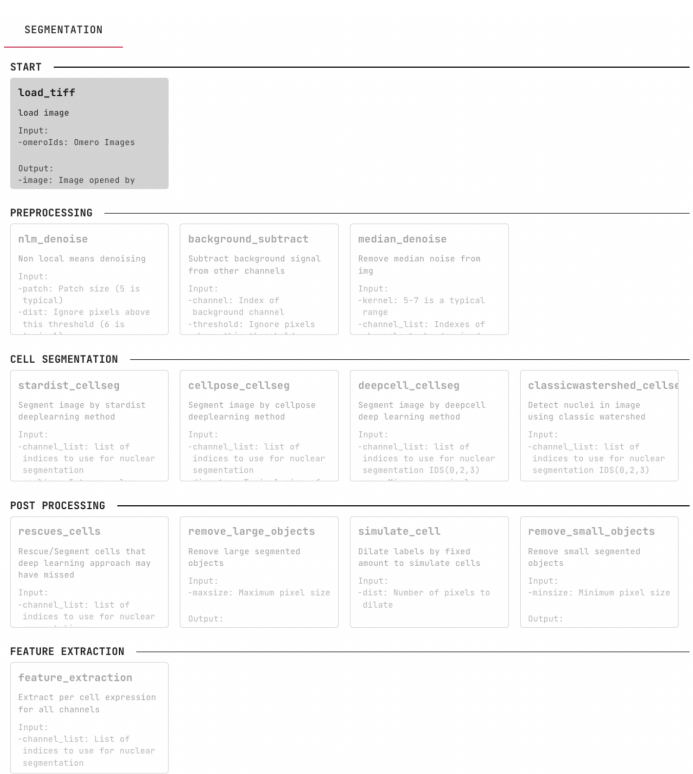
**Figure 6: SPEX identifies structural composition of tonsil tissue. A:** The histologic image is color-coded by 17 structural hierarchy categories identified by the marker expression pattern. For example, in germinal centers, green represents Ki-67$^{+}$, BCL2 low, CD45 low B-cells in the "dark zone;" these cluster in the upper sector of the UMAP germinal center lobe, and are shown as cluster 844 in **C**; hovering over the cluster 844 Ki-67 cell in **C** shows the average expression level to be 61.93 units. For comparison, the "light zone" germinal center B-cells are shown as red in A-C, and as cluster 275 in D; hovering over the Ki-67 cell for this cluster shows the average expression level to be 12.29. **D:** The spatial distribution and gene expression profiles of cells in each expression cluster can be explored by overlaying the relative expression of selected antigens on the tissue image and UMAP. Ki67 expression (left) is highest in proliferating squamous epithelium (cluster 417) and germinal center cells of the dark zone subset (cluster 844); CD68 expression (right) is highest in cells in the germinal center (cluster 85), paracortex (cluster 749) and squamous tissue (cluster 133). **E:** Cell populations that co-localize (red circles) with or segregate from (blue circles) other populations are displayed graphically; the size of the circles represents the magnitude of the result. Rows represent the neighborhood of a cell phenotype of interest and columns the enrichment or depletion of a cell in other neighborhoods. Cell populations are mapped to the tissue architecture, showing examples of avoidance: germinal center B-cells (cluster 228), red vs. CK+, TIM3+ squamous epithelium (cluster 512), light blue, and co-occurrence: CK+, TIM3high squamous epithelium (cluster 370), yellow vs. CD68+ PD-L1high macrophages (cluster 133), light blue.

# 6 Supporting Information

A


Analytical readouts (segmentation, phenotype, spatial relationships)
Graphical User Interface
S P E X
GNE Developed algorithms
External algorithms
OMERO
Centralized image database
High-dimensional imaging data

B


SEGMENTATION
START
load_tiff
load image
Input:
-omeroIds: Omero Images
Output:
-image: Image opened by
PREPROCESSING
nlm_denoise
Non local means denoising
Input:
-patch: Patch size (5 is typical)
-dist: Ignore pixels above this threshold (6 is
background_subtract
Subtract background signal from other channels
Input:
-channel: Index of background channel
-threshold: Ignore pixels
median_denoise
Remove median noise from img
Input:
-kernel: 5-7 is a typical range
-channel_list: Indexes of
CELL SEGMENTATION
stardist_cellseg
Segment image by stardist deeplearning method
Input:
-channel_list: list of indices to use for nuclear segmentation
cellpose_cellseg
Segment image by cellpose deeplearning method
Input:
-channel_list: list of indices to use for nuclear segmentation
deepcell_cellseg
Segment image by deepcell deep learning method
Input:
-channel_list: list of indices to use for nuclear segmentation IDS(0,2,3)
classicwastershed_cellse
Detect nuclei in image using classic watershed
Input:
-channel_list: list of indices to use for nuclear segmentation IDS(0,2,3)
POST PROCESSING
rescues_cells
Rescue/Segment cells that deep learning approach may have missed
Input:
-channel_list: list of indices to use for nuclear
remove_large_objects
Remove large segmented objects
Input:
-maxsize: Maximum pixel size
Output:
simulate_cell
Dilate labels by fixed amount to simulate cells
Input:
-dist: Number of pixels to dilate
remove_small_objects
Remove small segmented objects
Input:
-minsize: Minimum pixel size
Output:
FEATURE EXTRACTION
feature_extraction
Extract per cell expression for all channels
Input:
-channel_list: List of indices to use for nuclear segmentation

C

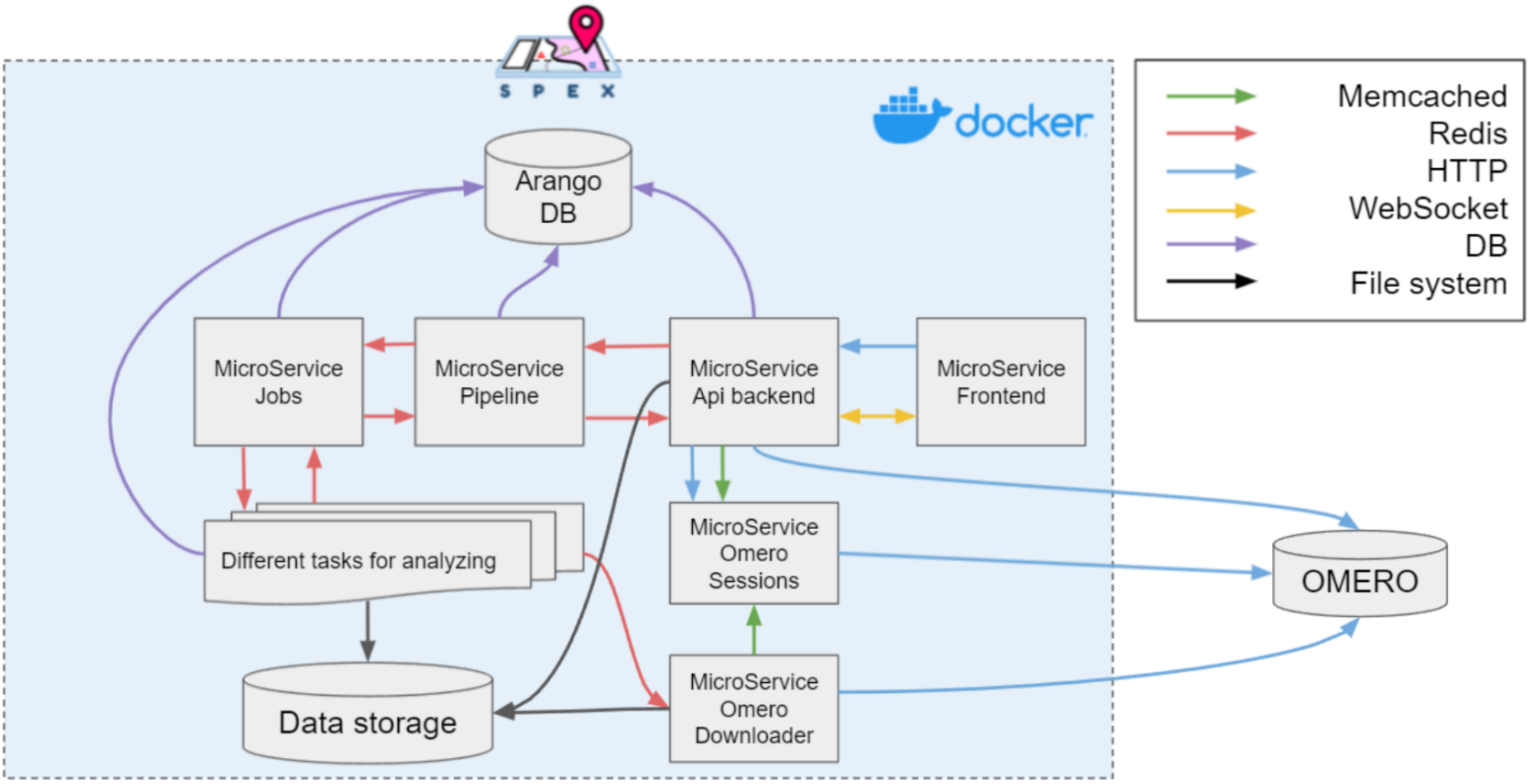

S P E X
docker
Arango DB
MicroService Jobs
MicroService Pipeline
MicroService Api backend
MicroService Frontend
Different tasks for analyzing
MicroService Omero Sessions
OMERO
Data storage
MicroService Omero Downloader
Memcached
Redis
HTTP
WebSocket
DB
File system

**Figure S1: SPEX framework and architecture to analyze and visualize spatially resolved omics data. A:**SPEX supports inputs from diverse spatial molecular technologies. It provides both infrastructure and quantitative analysis methods (segmentation, phenotyping, spatial relationships, etc) that allow for streamlined access to open source image data management systems along with preprocessing, analysis and interactive visualization of spatial expression data. **B:** SPEX image processing module. **C:** Architectural overview of the SPEX system.Redis as a data bus. Pub/Sub, LPOP/RPUSH/LPUSH (Queue) features provide a serialization and message ordering capability. ArangoDB as a permanent database. Data Storage (File Storage) - NFS mounted to a VM. Any technology that supports NFS can serve as Data Storage. Used for images, files generated as a result of calculations etc. Frontend microservice is an SPA on React that communicates with the Backend using HTTP requests to the JSON API. Web sockets are employed to promptly update the UI. API Backend defines the business logic for presenting data to the user, setting tasks, and gaining access to system resources. Omero Sessions handles communication with the OMERO server, manages sessions, utilizing Memcached for sessions storage and retrieval Omero Downloader service allows quick images downloading from OMERO. Pipeline microservice manages analysis pipelines, monitors their state, and queues them for an execution. Jobs microservice manages the analysis tasks, monitors their status and forms a task execution queue for a specific pipeline. Tasks service launches tasks (scripts), prepares incoming data and collects the execution result.

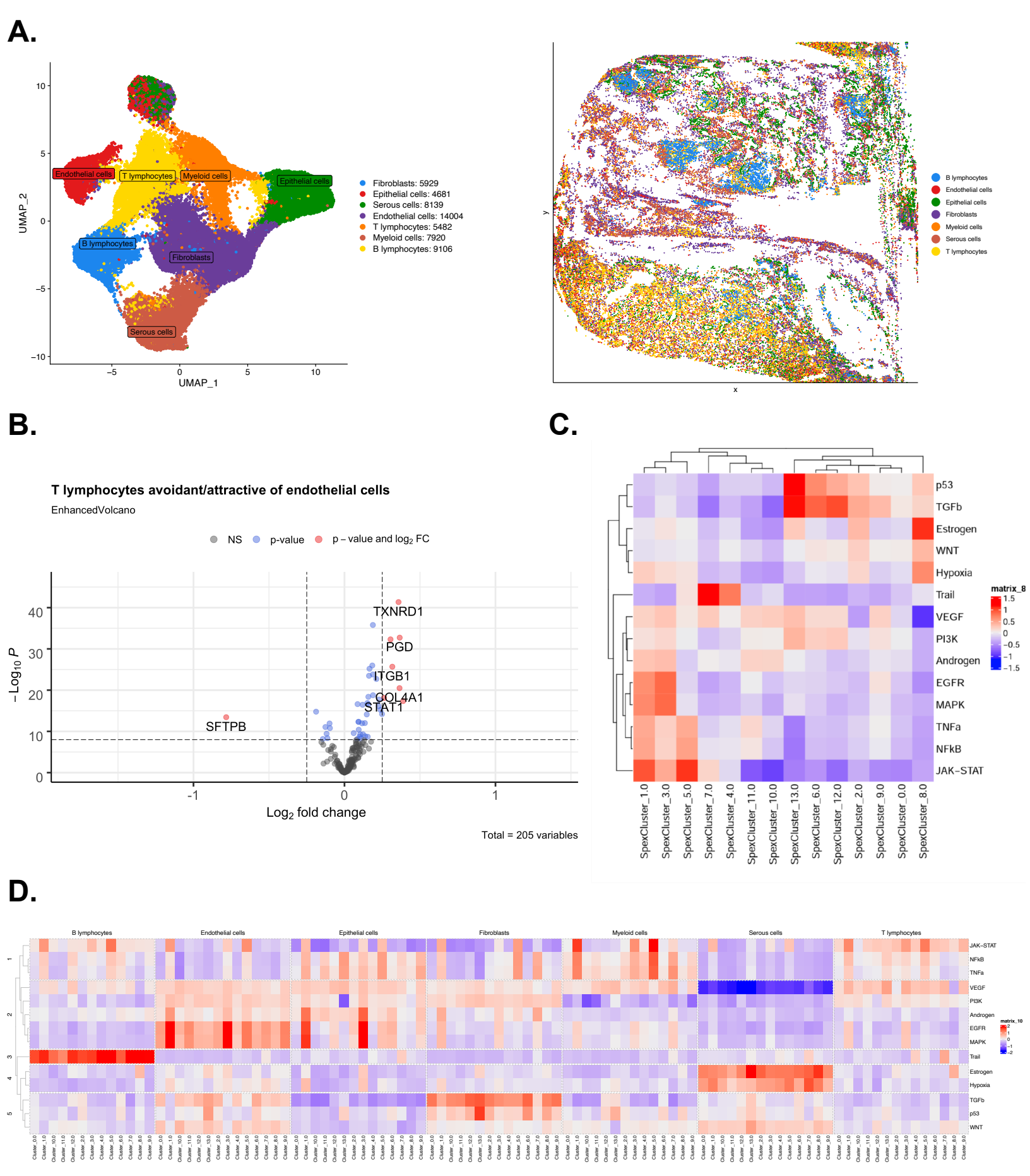
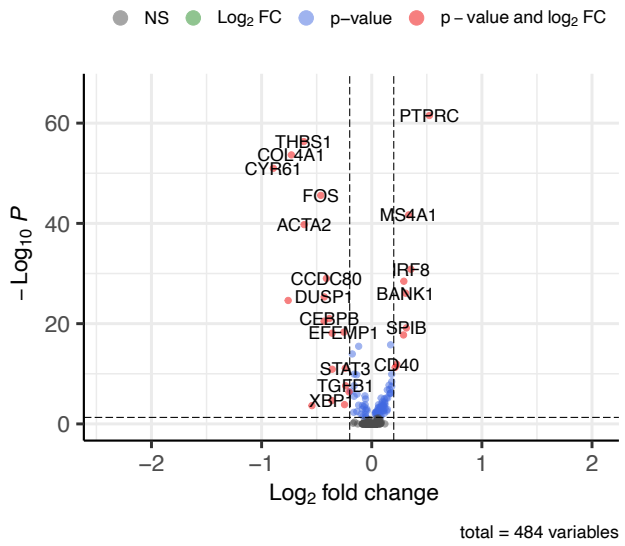
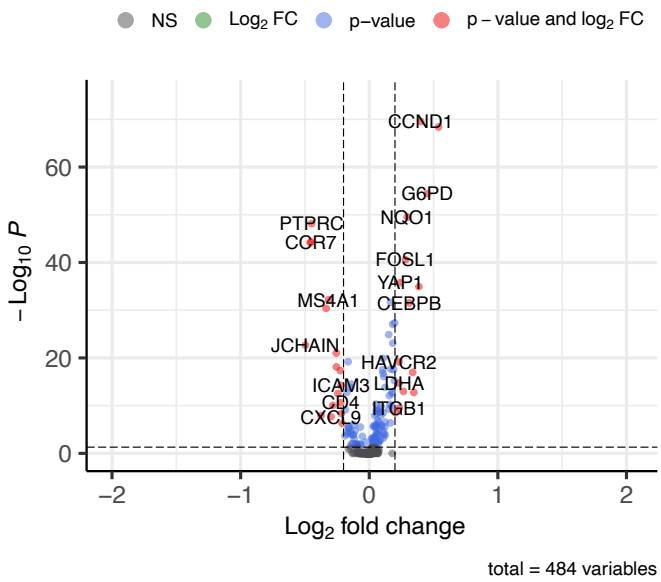

A.
Endothelial cells
T lymphocytes
Myeloid cells
Epithelial cells
B lymphocytes
Fibroblasts
Serous cells
UMAP_2
UMAP_1
Fibroblasts: 5929
Epithelial cells: 4681
Serous cells: 8139
Endothelial cells: 14004
T lymphocytes: 5482
Myeloid cells: 7920
B lymphocytes: 9106
B lymphocytes
Endothelial cells
Epithelial cells
Fibroblasts
Myeloid cells
Serous cells
T lymphocytes
B.
T lymphocytes avoidant/attractive of endothelial cells
EnhancedVolcano
NS
p-value
p – value and log2 FC
TXNRD1
PGD
ITGB1
COL4A1
STAT1
SFTPB
−Log10 P
Log2 fold change
Total = 205 variables
C.
p53
TGFb
Estrogen
WNT
Hypoxia
Trail
VEGF
PI3K
Androgen
EGFR
MAPK
TNFa
NFkB
JAK−STAT
matrix_8
SpexCluster_1.0
SpexCluster_3.0
SpexCluster_5.0
SpexCluster_7.0
SpexCluster_4.0
SpexCluster_11.0
SpexCluster_10.0
SpexCluster_13.0
SpexCluster_6.0
SpexCluster_12.0
SpexCluster_2.0
SpexCluster_9.0
SpexCluster_0.0
SpexCluster_8.0
D.
B lymphocytes
Endothelial cells
Epithelial cells
Fibroblasts
Myeloid cells
Serous cells
T lymphocytes
JAK−STAT
NFkB
TNFa
VEGF
PI3K
Androgen
EGFR
MAPK
Trail
Estrogen
Hypoxia
TGFb
p53
WNT
matrix_10

**Figure S2: SPEX Spatial analytics and spatially informed pathway analysis. A:** UMAP depicting identified cell types using MERFISH and their location in the slide of a human lung tumor sample. **B:** Volcano plot for the differentially expressed genes between the T cells avoidant of endothelial cells and those that are attractive. **C:**. SPEX communities exhibit characteristic pathway activities. **D:** Pathway activities computed in each cell type within each SPEX community.

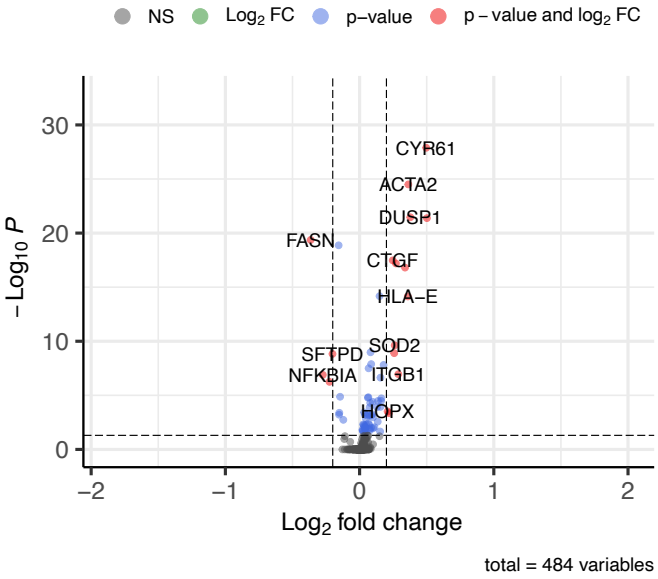
B lymphocytes_SpexCluster_7.0vsB lymphocytes_SpexCluster_4.0



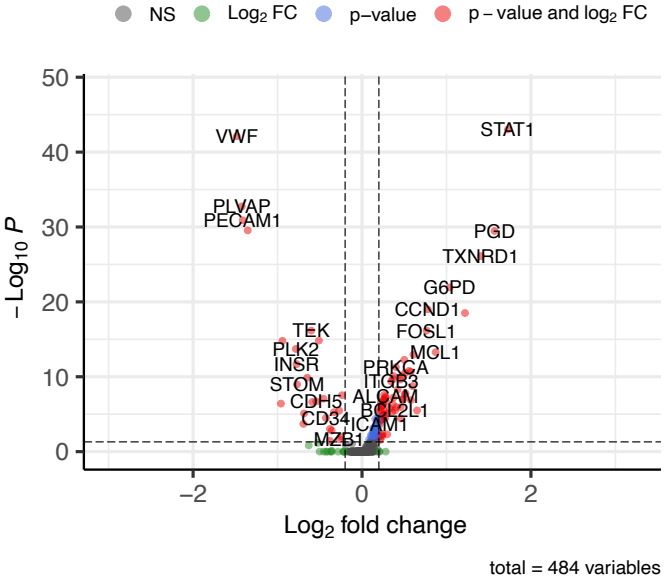

T lymphocytes_SpexCluster_1.0vsT lymphocytes_SpexCluster_5.0



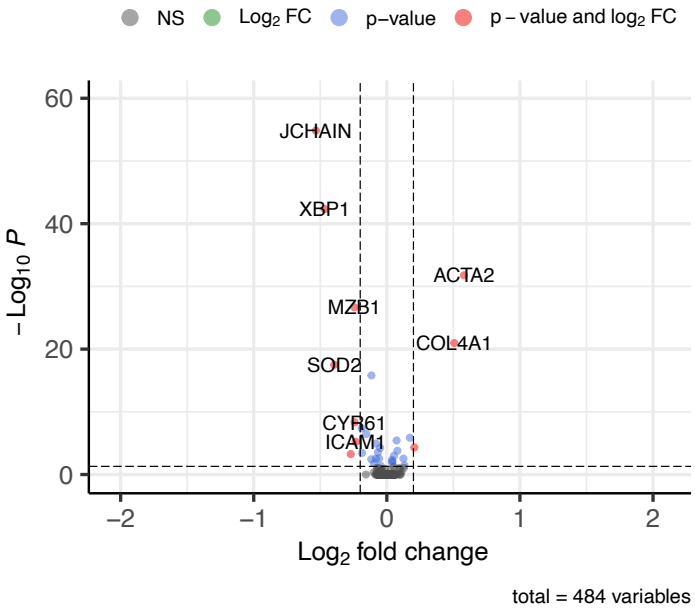

Epithelial cells_SpexCluster_0.0vsEpithelial cells_SpexCluster_10.0




Endothelial cells_SpexCluster_1.0vsEndothelial cells_SpexCluster_8.0




Myeloid cells_SpexCluster_1.0vsMyeloid cellss_SpexCluster_3.0




Serous cells_SpexCluster_0.0vsSerous cells_SpexCluster_8.0

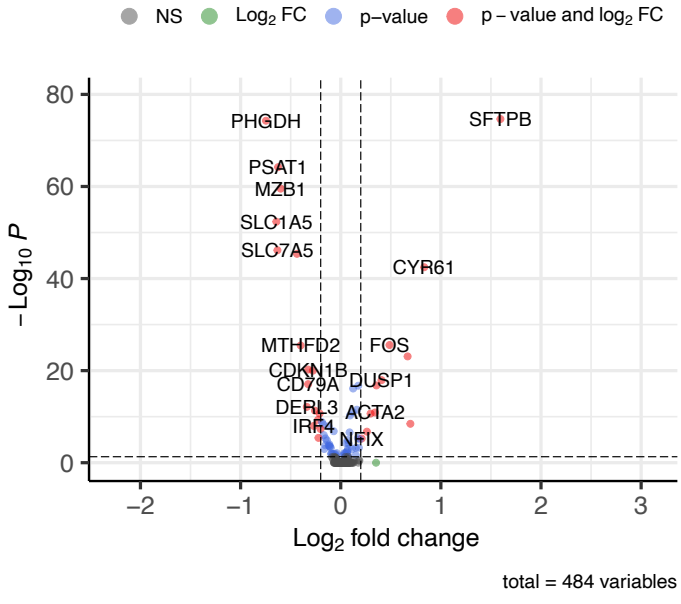
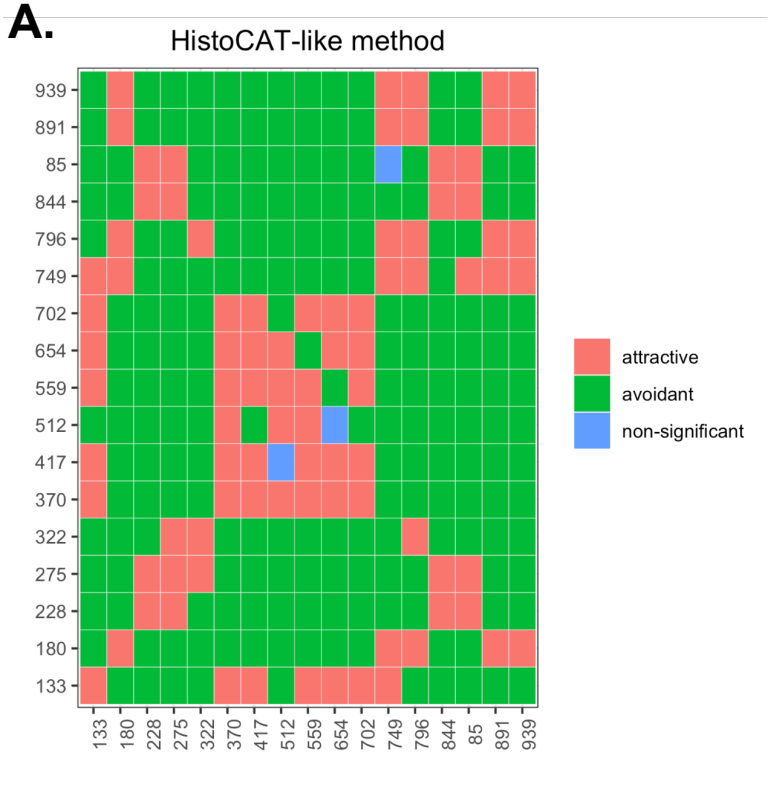
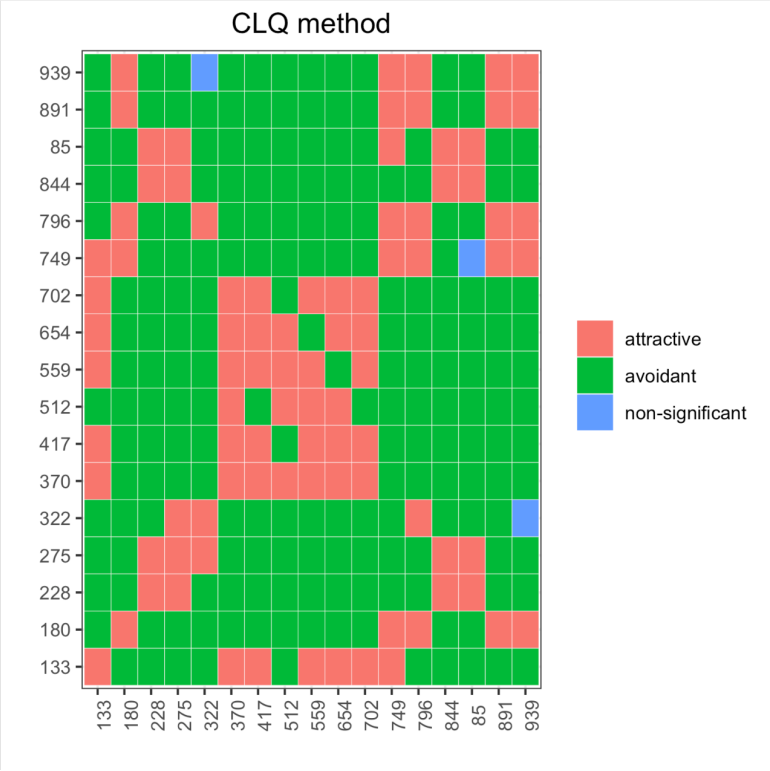


**Figure S3: SPEX Spatial analytics and spatially informed pathway analysis.** Volcano plots demonstrating differential expression in several cell types within the context of distinct SPEX communities.

**A.**




**B.**

control_tonsil_Cluster IDs

| Cluster | Zone | Classification | Expression Pattern |
|---|---|---|---|
| **180** | Germinal Center | Stromal fibroblasts | Ki67+/-, Collagen+ |
| **844** | Germinal Center | GC "Dark Zone" | Ki67+/-, BCL2-, CD45 low |
| **228** | Germinal Center | GC Transition | Ki67+/low, BCL2-, CD45 low |
| **275** | Germinal Center | GC "Light Zone" | Ki67 low, BCL2 low, CD45 low |
| **85** | Germinal Center | GC macrophages | CD68+, CD20 (by contamination?) |
| **749** | Mantle Zone and Paracortex | Macrophages | CD68+ |
| **891** | Mantle Zone and Paracortex | Stromal fibroblasts and endothelium | CD31+/-, Collagen+ |
| **939** | Mantle Zone and Paracortex | T Cells | BCL2+, CD8+, CD20 low |
| **796** | Mantle Zone and Paracortex | Mantle zone | BCL2+, CD20+ |
| **322** | Mantle Zone and Paracortex | GC "cap" | BCL2+, CD20+, CD8- |
| **512** | Epithelium | Squamous epithelium | CK+, TIM3+ |
| **370** | Epithelium | Squamous epithelium | CK+, TIM3 high |
| **559** | Epithelium | Vascular Endothelium | CD31+, PD-L1+ |
| **702** | Epithelium | Squamous epithelium | CK+, CD20 high |
| **417** | Epithelium | Proliferating epithelial cells | CK+, Ki67+ |
| **654** | Epithelium | Lymphoepithelial | CD68+ or CD8+, CK high |
| **133** | Epithelium | Macrophages | CD68+, PD-L1 high |

**Figure S4: Results comparison of HistoCAT-like vs CLQ methods. A:** Each cell in the heatmap represents the cell-cell co-occurrence conclusion of the cell phenotype pair. Red color indicates the pair is significantly attractive; blue color means significantly avoidant; green color implies non-significance. Overall we have 283 concordant pairs between these two methods among all 289 pairs. **B:** Cluster identities.

**Table S1:** Antibodies, conjugates, clones and working concentrations used in tonsil multiplex staining

| Antibody | Conjugate | Clone | Concentration ($\mu$g/mL) |
|---|---|---|---|
| CD20 | 161-Dy | H1 | 3 |
| CD45 | 152-Sm | D9M8I | 7 |
| CD68 | 159-Tb | KP1 | 0.1 |
| CD8 | 162-Dy | CD8/144B | 2 |
| Collagen type I | 169-Tm | Polyclonal | 1 |
| Cytokeratin | 176-Yb | AE1/AE3 | 0.8 |
| Histone H3 | 171-Yb | D1H2 | 0.5 |
| Ki67 | 168-Er | B56 | 3 |
| BCL2 | 146-Nd | EPR17509 | 7 |
| CD25 | 175-Lu | EPR6452 | 5 |
| CD31 | 151-Eu | EPR3094 | 5 |
| CD336/Tim3 | 154-Sm | D5D5R | 7 |
| PD1 | 165-Ho | EPR4877 | 5 |
| PDL1 | 150-Nd | 73-10 | 10 |

**Table S2:** Inconsistent pairs of the human tonsil tissue between the histoCAT and CLQ methods.

| From type | To type | histoCat-like conclusion | CLQ conclusion | Cell population of from type | Cell population of to type |
|---|---|---|---|---|---|
| 0_607 | 0_180 | non-significant | attractive | 1172 | 719 |
| 0_939 | 0_322 | avoidant | non-significant | 1107 | 2772 |
| 0_512 | 0_417 | non-significant | avoidant | 1387 | 1035 |
| 0_180 | 0_465 | non-significant | attractive | 719 | 211 |
| 0_891 | 0_465 | non-significant | attractive | 1892 | 211 |
| 0_654 | 0_512 | non-significant | attractive | 1460 | 1387 |
| 0_180 | 0_607 | non-significant | attractive | 719 | 1172 |
| 0_465 | 0_654 | non-significant | avoidant | 211 | 1460 |
| 0_465 | 0_749 | non-significant | avoidant | 211 | 1570 |
| 0_85 | 0_749 | attractive | non-significant | 655 | 1570 |
| 0_749 | 0_85 | non-significant | attractive | 1570 | 655 |
| 0_322 | 0_939 | avoidant | non-significant | 2772 | 1107 |