scientific reports

Check for updates

OPEN

Automated spatial omics landscape analysis approach reveals novel tissue architectures in ulcerative colitis

Derek R. Holman[1,5]✉, Samuel J. S. Rubin[1,5], Mariusz Ferenc[2,5], Elizabeth A. Holman[1,5], Alexander N. Koron[1], Robel Daniel[1], Brigid S. Boland[3], Garry P. Nolan[4], John T. Chang[3] & Stephan Rogalla[1]✉

The utility of spatial omics in leveraging cellular interactions in normal and diseased states for precision medicine is hampered by a lack of strategies for matching disease states with spatial heterogeneity-guided cellular annotations. Here we use a spatial context-dependent approach that matches spatial pattern detection to cell annotation. Using this approach in existing datasets from ulcerative colitis patient colonic biopsies, we identified architectural complexities and associated difficult-to-detect rare cell types in ulcerative colitis germinal-center B cell follicles. Our approach deepens our understanding of health and disease pathogenesis, illustrates a strategy for automating nested architecture detection for highly multiplexed spatial biology data, and informs precision diagnosis and therapeutic strategies.

Keywords Spatial omics, Computational analysis, Inflammatory bowel disease

Ulcerative colitis (UC) is a gastrointestinal immune-mediated disease that occurs due to a combination of environmental stressors in a genetically predisposed individual. Endoscopically and histologically, UC is characterized by visible erythema, friability, and ulcers with inflammatory infiltrate. The clinical pattern of the disease itself presents with highly variable phenotypes, disease progression, and responsiveness to different therapies. Therapies targeting inflammation are modestly effective, but rates of primary non-response (13–40%) and loss of response (46% of the remainder) are relatively high[1,2]. Response rates could potentially be improved by more precise UC molecular subtype stratification[3,4].

Recent insights from spatial omics demonstrate that emergent properties from changes in tissue structural modules called architectures[5,6] are associated with disease subtypes, therapeutic response, and disease pathogenesis. This suggests that cell-type and tissue architectures serve as modular, functional building blocks. However, little is known about how changes in tissue structures are linked to disease initiation, progression, and therapeutic response. Improving our understanding of the role that architectural changes play in disease progression and recovery can enhance precision diagnostics and therapy by identifying potential functional relationships between patient therapy outcomes and their underlying tissue architectures. To accurately identify disease-associated cell-types and their relevant architectures, one must appropriately select the level of granularity for their cell annotations[7].

Spatial omics technologies historically applied to UC cell annotation were developed using analysis pipelines modified from dissociated single-cell analyses that lack explicit spatial information[8]. This critical step was determined by dimensional reduction and clustering, automatic cell-type identification based on a labeled reference from publicly available databases or datasets, or supervised learning[9]. Spatial considerations were only incorporated while validating cell identity by directly displaying those annotations on the multiplexed fluorescent image or when interrogating the spatial distributions of functional markers. Recent advances build on these historical

[1]Division of Gastroenterology and Hepatology, Department of Medicine, Stanford University, Stanford, CA, USA. [2]Nalecz Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences, Warsaw, Poland. [3]Division of Gastroenterology, Department of Medicine, University of California San Diego, San Diego, CA, USA. [4]Department of Pathology, Stanford University, Stanford, CA, USA. [5]These authors contributed equally: Derek R. Holman, Samuel J. S. Rubin, Mariusz Ferenc and Elizabeth A. Holman. ✉email: drholman@stanford.edu; srogalla@stanford.edu

approaches by incorporating spatial information such as neighboring cell types[10]. Their results, which concluded that classification accuracy is increased by taking neighboring cell types into account, suggest that biological architectures and spatial compartmentalization may also be promising candidates for refining cellular annotation.

Another critical challenge in cell-type annotation arises from marker intensities that are not readily thresholded when presented in a dissociated-cell format[11]. Approaches that incorporate spatial considerations at the cellular level can address this challenge. By using patterns of spatial heterogeneities as a guide, marker intensities are evaluated in the context of specific, local architectures rather than the tissue as a whole. However, these new methods rely on deep learning technologies coupled with novel cell-type detection capabilities and may rely on manually annotated datasets[7]. Depending on the specific experimental or clinical question and because of its reliance on trained datasets, this approach may not adequately address the multiple levels of cell labeling details, also known as annotation granularity[7], necessary for a deeper understanding of a dataset across multiple spatial scales.

New technologies to identify and analyze architectures in tissues are crucial for a deeper understanding of the increasingly abundant publicly available spatial omics datasets. Many similar approaches have been pioneered in the geospatial sciences, including ecology and landscape ecology, as reviewed in Newman et al.[12]. Just as recent advances in spatial imaging technologies require the development of analytical tools and strategies for identifying, characterizing, and understanding tissue architectures and cellular interplay[5,8,10,13], the development and proliferation of aerial photography in the 1930s inspired the development of tools for identifying and characterizing landscapes at multiple scales. These exploratory landscape analysis techniques offer insight into complex systems containing many spatially-resolved interacting components that interact across multiple scales to produce emergent properties. One critical parameter used in landscape analysis to represent the complexity of a system is physical entropy, which represents the spatially-resolved disorder in a system[12].

Since tissue architectures by definition represent compartmentalized function and chemistry, tissue architectures are therefore regions of decreased local entropy, or disorder. We anticipate that the quantification and ordering of multi-scale phenomena as addressed by hierarchical patch dynamics[14,15] and their emergent properties[16] will be widely used in spatial biology, especially once extended to multivariate parameter screens. Here we perform a proof-of-principle implementation of spatial metrics for the purpose of demonstrating the method's applicability. We use spatial analysis by distance indices (SADIE) for pattern-dependent feature extraction based on observed randomness, regularity, or aggregation[17]. We generate both global and local metrics of randomness, regularity, and aggregation, allowing for architecture detection and visualization by thresholding as well as minimizing the need for assumptions regarding the spatial scale of cellular interactions. Here SADIE iteratively adjusts two-dimensional maps containing the x,y coordinates of individuals–in this case thresholded marker-positive cells, until the adjusted spatial distribution nears regularity. SADIE then assigns both global and local indices of aggregation summarizing the total distance individuals move[17–19]. We envision this strategy will be most effective for automated architecture-guided cell-type annotation, and should be implemented in formal workflows as a user-tuned add-on module that refines existing, rapid annotations by more generalized automated approaches (Fig. 1A)[10,13].

Beyond label-transfer, we implement additional strategies for integrating spatially-resolved single cell technologies with existing single cell datasets. We integrate existing UC datasets from single-cell RNA sequencing (scRNA-seq), cytometry by Time of Flight (CyTOF), and Co-detection by Indexing (CODEX) datasets available in literature. The integration of these three technologies (scRNAseq, CyTOF and CODEX), each of which has distinctive yet complementary strengths and weaknesses (Fig. 1B), gives additional certainty to conclusions drawn from any one technology. Of the three technologies, only CODEX explicitly addresses spatial resolution. As shown by Zhao et al.[20], single tissue sections from a biopsy are unlikely to be statistically representative of that biopsy at scales ranging from the cellular to tissue levels. This consideration dominates when interrogating rare cell-types, rare architectures, and large but spatially complex architectures. Unlike CODEX, both scRNA-seq and CyTOF dissociate tissues into single cells and perform a random sampling prior to data acquisition. Aside from known differential cell sensitivity to the dissociation process, both scRNA-seq and CyTOF yield data that are statistically representative of the original biopsy[21,22].

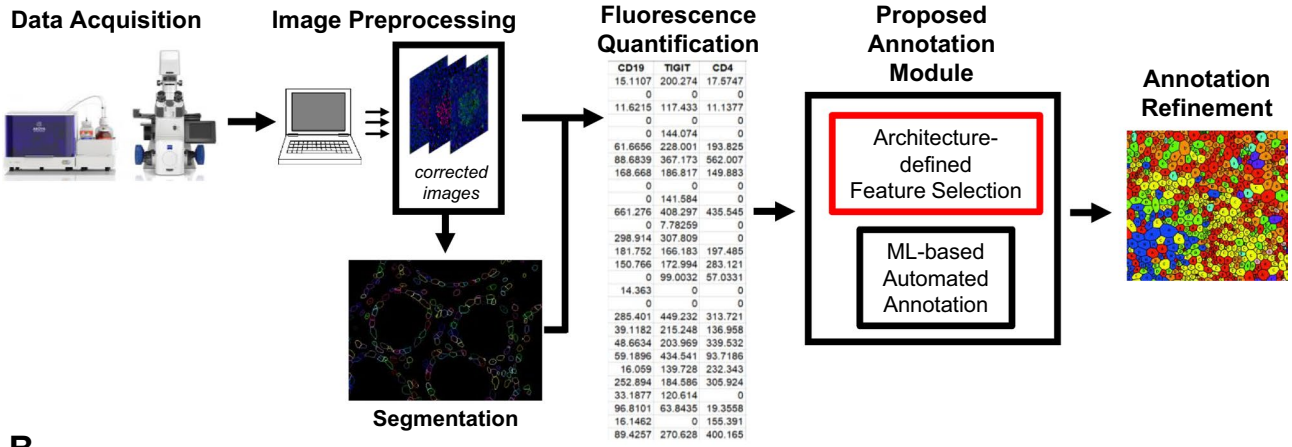
CyTOF observations are more similar to CODEX in that both examine functional protein-level data while scRNA-seq provides regulatory transcript-level data. Integrating CyTOF and CODEX datasets can be challenging because of dimensional mismatch in the choice of observed parameters. Generally, there are relatively few (<100) observed dimensions in comparison to scRNA-seq. These parameters are selected as an a priori panel. When datasets from different investigators are not specifically constructed to complement each other, the probability of marker overlap is low outside of core canonical cell markers[23,24]. In contrast to CyTOF and CODEX, scRNA-seq datasets often contain an order of magnitude (>2000) greater observed dimensions. The probability of marker overlap is therefore increased. The well-known bias towards high-abundance transcripts in scRNA-seq poses an additional challenge[25]. We develop and demonstrate a workflow strategy that makes use of surrogate observations and superordinate architectures coupled with spatially-resolved per-architecture analysis rather than per-patient analysis to overcome the aforementioned dimensional mismatch challenges.

Results

Datasets

We used three publicly available single-cell resolution UC datasets from different tissue blocks–CODEX, CyTOF, and scRNA-seq–to demonstrate the effectiveness of our approach (Supplemental Table S1). The CODEX dataset was derived from colonic biopsies from 24 UC patients and 8 healthy HC patients[23]. The scRNA-seq dataset was derived from 10 HC and 10 UC patients, with samples from both tissue biopsies and patient blood[26]. The CyTOF dataset includes peripheral blood mononuclear cells (PBMCs) from 20 UC and 12 matched HC patients, as well as mononuclear cells from paired blood and colon tissue biopsy samples from an additional 12 UC patients[24].

A



B

Technology	Spatial Resolution	Representative	Regulation	Functional	Maximum Number of Markers
CODEX	*Yes*	*No*	*No*	*Yes*	*< 100*
scRNA-seq	*No*	*Yes*	*Yes*	*No*	*> 1000*
CyTOF	*No*	*Yes*	*No*	*Yes*	*< 100*

Figure 1. Proposed Implementation and Datasets. We propose to implement our add-on as a supplementary annotation module for existing annotation methods, including rapid, high-throughput ML methods (**A**). Characteristics the of different single-cell technologies datasets used here (**B**).

Blood samples were classified as being from HC, UC-flare, or UC-remission. Inflamed biopsy samples were further matched to uninflamed control tissues from the same patient.

Workflow

Our analytical workflow supplements canonical cell-type markers with B-cell follicle architecture-dependent features. We define a tissue landscape as "an area that is spatially heterogeneous in at least one factor of interest"[27]. Thus, our approach identifies marker-dependent spatial heterogeneities in cellular level representations within B-cell follicles. Architecture-dependent features are identified through identifying marker-positive cells followed by an assessment of aggregation value (I_a) using SADIE[17]. Briefly, in this case, I_a is a representation of the distance from regularity of a collection of individuals. I_a values exceeding 1 indicate aggregation, I_a values close to 1 are indicate randomness, while values less than 1 indicate regularity[17]. Once key cell populations are identified, we perform a secondary parameter screen for markers representing cell types that are indicative of architectures and their associated lower-granularity cell annotations. We assign increased weights to markers with statistically significant ($p < 0.05$) aggregation values (I_a) exceeding 1, that are associated with only a small number of marker-positive cells (Supplemental Table S2), for clustering purposes.

By visualizing these screened parameters using scRNA-seq and CyTOF, we are able to characterize biological differences between UC and control (Fig. 2A).

CODEX automated feature selection identifies complex architectures contributing to UC germinal center B cell follicles

In order to simplify our analysis pipeline, we only selected cells that lay within follicle-associated regions of interest (RoIs) as determined by CD21 + cell aggregates visible in fluorescent images. Altogether, this process resulted in a total of 4 control and 51 UC-associated follicle RoIs (Supplemental Table S3), one of which is displayed in Fig. 2B.

Of a total of 53 fluorescent marker parameters, 13 were selected for cell identification based on canonical cell type markers, marker staining quality, and spatial distribution within representations of the follicles as assessed by SADIE. Spatial-associated parameters were selected based on their I_a values and statistical significance, weighted by the fraction of cells above the designated marker intensity value (Supplemental Table S1). By directly incorporating spatial coordinates and sequential sub-clustering steps into the clustering workflow followed by annotation refinement (Fig. 2C), we identify 10 different architecturally-relevant cell annotations (Fig. 2D) (see Methods), of which two are of particular interest. The first of these identities is CD56 + B cells which also express elevated levels of the established proliferation marker Ki67. This suggests that the CD56 marker on B-cells is either directly associated with proliferation or early stages of peripheral B-cell maturation. The second

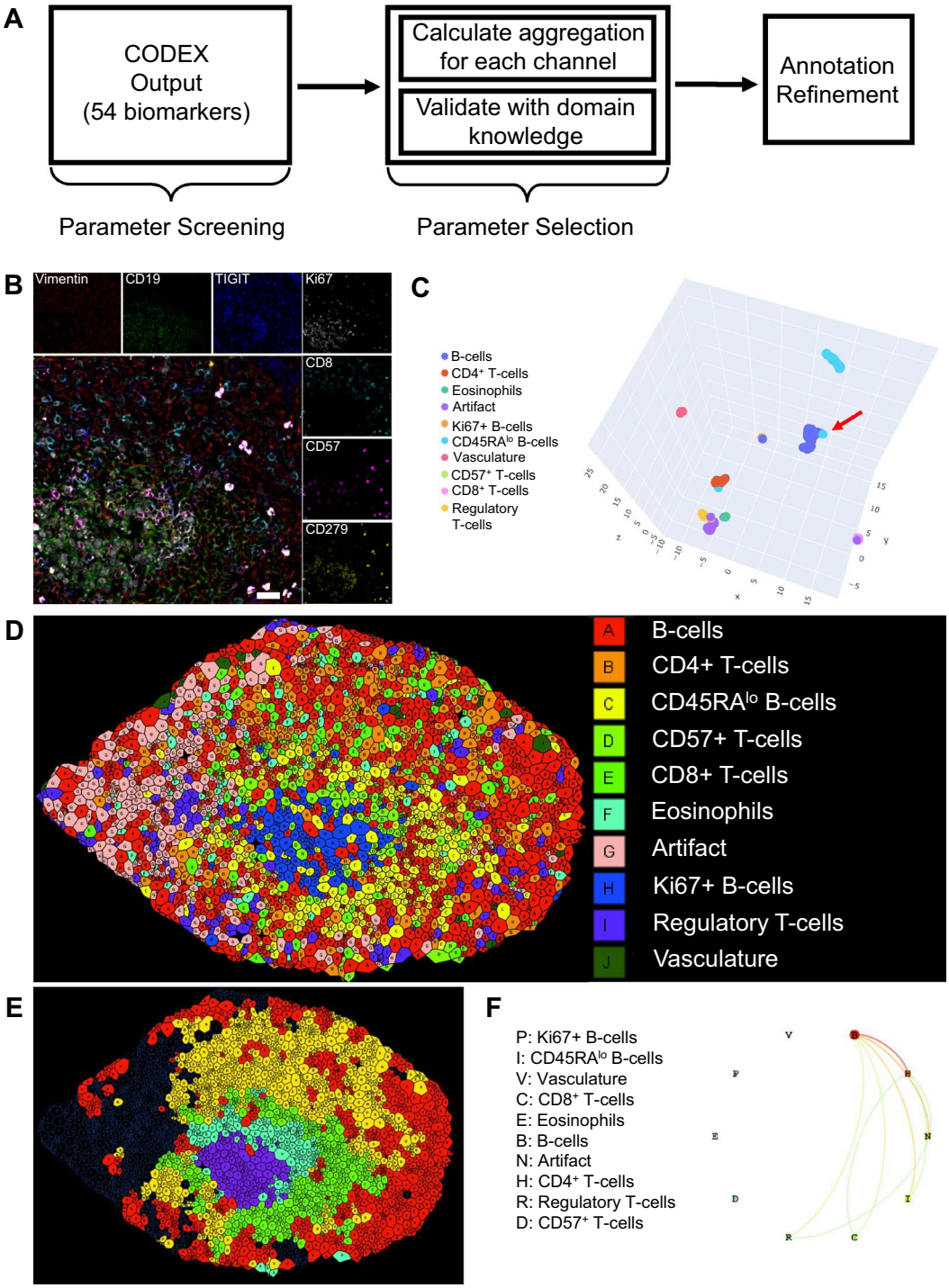


Figure 2. SADIE Analysis Identifies Rare Cell Types and Follicular Architectures. SADIE-guided workflow for cell annotation refinement, in which channels are initially screened in a quality control step. Aggregation values (I_a) are then obtained for each channel in the architecture of interest. High I_a-value channels are selected and supplemented with known canonical markers (parameter selection) for annotation refinement (**A**). An example 7-channel fluorescent image of a proliferating UC B-cell follicle (**B**). Scale bar = 300 μm. Note the honeycomb-like profile of membrane markers that results in segmentation difficulties and "bleed" into adjacent cells. Nonetheless, by directly incorporating spatial coordinates for further annotation refinement, we are able to re-annotate cells that appear to have been mis-clustered ((**C**), red arrow). Data space is visualized using a UMAP reduction (**C**). The result of SADIE-guided parameter selection and subsequent annotation refinement is displayed on the associated Voronoi image, demonstrating the successful annotation even of rare cell types (**D**). Associated cellular neighborhoods are displayed in their Voronoi representation (**E**). Pairwise interactions, in which the size of each circle represents the fraction of all cells and the width of the line represents the thresholded ratio of interactions (**F**).

of these identities is a, to our knowledge previously unreported, rare cell population of CD57 + CD279 + CD4 + T cells (Fig. 2D) (Supplemental Fig. S1) that lies exclusively within the proliferating center of germinal-center B cell follicles. We therefore select CD56 and CD57, Ki67-like proliferation markers (such as PCNA), and established markers of T-cell activation and exhaustion like CD279 for downstream validation in the representative technologies scRNA-seq and CyTOF, to support our CODEX observations. Although the rarity of these cells renders them difficult to detect using standard dissociated-cells single cell omics technologies analysis platforms–they comprise fewer than 1 in 50,000 total cells–the spatial resolution of CODEX renders them easily identifiable.

We then performed neighborhood and pairwise analysis to determine relevant architectural 'zones' within our follicle. Six such zones were determined (Fig. 2F), though one was discarded due to artifact proximity. The innermost zone, zone 1, is characterized by a core of proliferating Ki67 + CD56 + B cells. Zone 2 is similar to zone 1, but is interspersed with CD4 + CD279 + T cells, a small minority of which also express CD57. Zone 3 expresses higher levels of CD45RA relative to zone 1 and begins to lose CD19 expression while retaining CD21 expression. Zone 4 exhibits reduced numbers of CD4 + CD279 + T cells, and we further begin to observe the presence of eosinophils and Foxp3 + T regulatory cells. Finally, zone 5 is marked by the encroachment of CD8 T cells. Together, these results suggest that germinal center B cell follicles in UC are highly complex architectural structures with numerous zones of development and maturation that are regulated by numerous different classes of T cells.

scRNA-seq verifies changes in proliferating B cells and associated T cell phenotypes

Due to the low number of visibly (Ki67 +) germinal center B cell follicles in the CODEX dataset relative to the number of tissue sections, we confirmed the increased presence of germinal center B cell follicles and potential follicular T cell subsets using scRNA-seq. When visualized using UMAP using gold standard approaches[28,29] (Supplemental Fig. S2), we are able to identify clusters corresponding to visually evident topographical features (Fig. 3A). Broadly, cells readily separate into three primary groups: plasma cells, T cells, and B cells (Fig. 3B). We then proceeded to map our candidate CODEX markers onto our scRNA-seq dataset visualization.

Critically, neither of the transcripts associated with candidate markers Ki67 and CD57 were present in the scRNA-seq dataset. Instead we used cell division-associated PCNA as a surrogate marker for CD56/Ki67, and PDCD1 (CD279) as a marker for the parent cell population of our CODEX CD4 + CD278/279 + CD57 + population (Fig. 3C). When visualized using UMAP topography, none of these markers were well-captured in their own clusters nor during subsequent subclustering. This indicates that gold-standard scRNA-seq analysis methods would have failed to identify these populations as being different between UC and control. This consequence reflects both limitations on data analysis as well as sequencing depth. However, because these markers are indicated as being spatially-segregated components, we are able to selectively visualize them in our scRNA-seq dataspace. We can therefore conclude that these cell types are increased, in the case of the CD56 + /Ki67 + B cell population, and that the permissive environment of parent cell populations is increased, in the case of the CD4 + CD278/279 + CD57 + T cell population, in UC patients.

Although the lack of markers Ki67 and CD57 point to concerns regarding the complete capture of cellular landscapes, our findings were further verified by an additional scRNA-seq dataset published after our original analysis (Supplemental Fig. S3)[30].

Mass cytometry verifies mucosal immune dysregulation in UC

When using CODEX, we observed that germinal-center B cell follicles predominantly existed in tissue sections from inflamed biopsies, although the proliferative center was only visible in a small subset of sections. Our scRNA-seq analysis confirmed the elevation of these architectures in UC versus healthy control biopsies. Similarly, when examining the CyTOF dataset, substantial numbers of CD56 + B-cells were observed in UC patients indicating the presence of germinal center follicles, even though direct proliferation-associated markers were not included in the panel. Unexpectedly, these cells were even more abundant in adjacent, uninflamed biopsies from UC patients (Fig. 3D). This appears to suggest that adjacent uninflamed biopsies are fundamentally different from biopsies from healthy controls and may retain disease-associated B-cell dependent architectural motifs. Exhausted CD4 + T-cells, not present in appreciable numbers in healthy controls, were similarly observed to be present in both inflamed and uninflamed UC biopsies. While the depletion of CD56 + cells from blood is observed during UC flare (Fig. 3E), suggesting that flare may involve additional trafficking from the blood to local sites of inflammation, these cell types do not appear to be represented by a dedicated cluster in our scRNA-seq analysis of blood samples as they were in our analysis of colonic biopsies (Supplemental Fig. S4). Furthermore, these cells lack proliferation-associated markers.

Discussion

Spatial omics hold great promise for precision medicine due to the possibility of linking multiscale cellular interactions to aggregate functions that may underlie disease heterogeneity. However, architecture detection is sensitive to cell annotations, especially with respect to annotation granularity, and there is a lack of methods that directly incorporate spatial heterogeneities in cell-type annotation. Here we show in Ulcerative Colitis (UC) that combining computational landscape ecology methodologies with spatial omics enables the ability to extract rare cell types. This approach is especially useful for cells that occupy restricted spaces within larger architectures such as germinal-center B-cell follicles that are likely to be of importance but would be otherwise missed due to their spatial distribution. We developed an automated pipeline to optimize existing cell annotation strategies for the characterization of tissue architectures. We demonstrate its success for increasing confidence in imaging results and bypass random-sampling weaknesses through the incorporation of existing, publicly available data from other single-cell omics technologies.

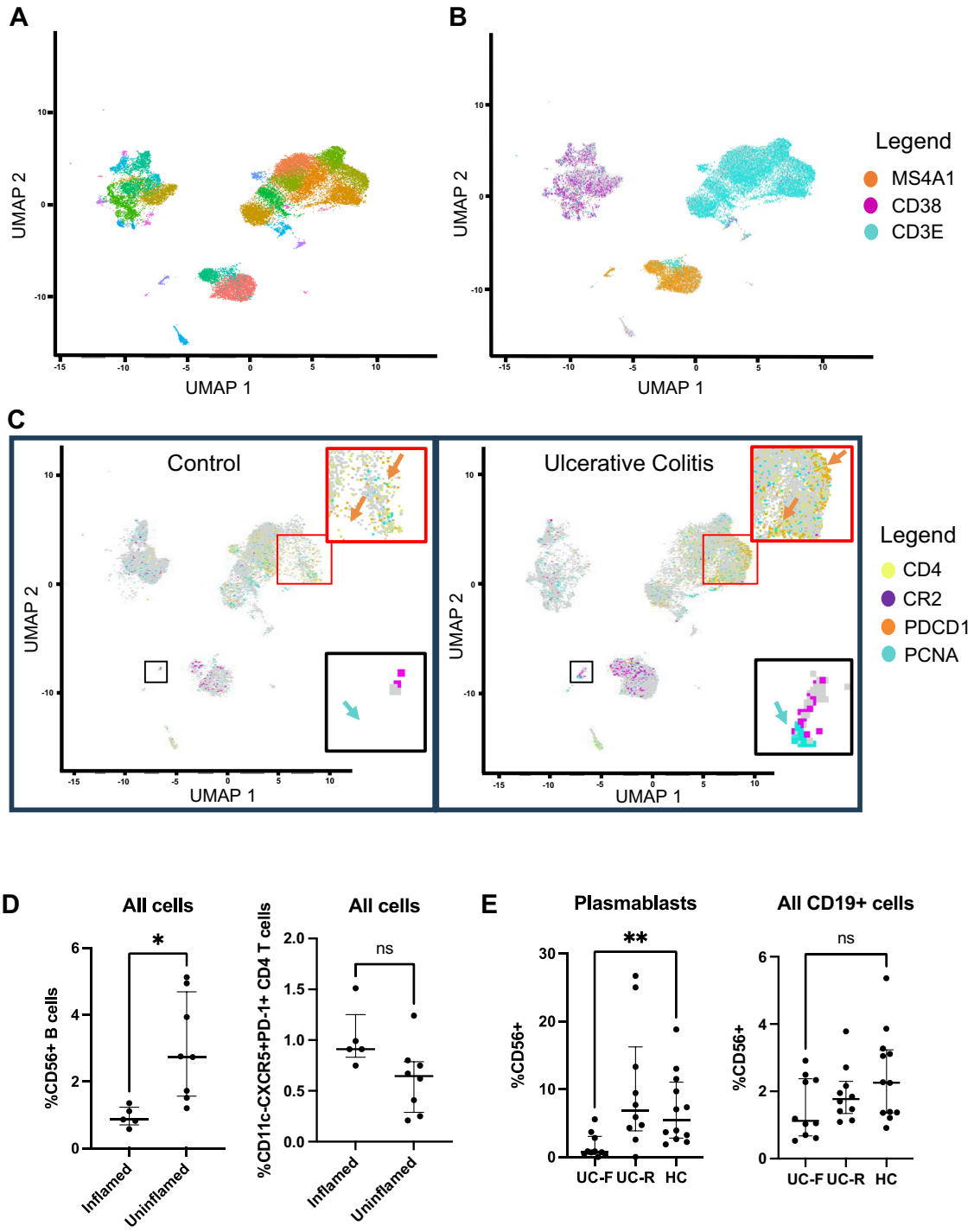


Figure 3. Evaluation of CODEX Results using scRNA-seq and CyTOF. Initial clustering results for scRNA-seq of colonic biopsies (**A**). UMAP representation and visualization of canonical markers identifies primary topographical features of interest: MS4A1 is a B-cell marker, CD3E is a T-cell marker, CD38 is a plasma cell marker (**B**). CD4 and PDCD1 (CD279) are displayed in the T-cell cluster ((**C**), red inset). Orange arrows indicate the PDCD1 + CD4 + T-cell region in UC samples ((**C**), red inset, right), as well as the corresponding depleted region in healthy controls (3C, red inset, left). CR2 (CD21) and PCNA (Ki67 surrogate) are displayed in the proliferating B-cell cluster ((**C**), black inset). Blue arrows indicate the PDNA + area of the MS4A1 + B-cell island in UC samples (3C, black inset, right) as well as the corresponding depleted region in healthy controls ((**C**), black inset, left). CyTOF reveals CD56 + B-cells and exhausted CD4 T-cells in UC patient colonic biopsies, though adjacent uninflamed controls have elevated proliferating B-cells (**D**). CD56 + B-cells as a fraction of all plasmablasts are reduced in blood during disease flare, relative to remission and healthy control (**E**), though similar differences are not observed as a fraction of all CD19 + cells.

Our method enables spatial feature-dependent annotations that were previously co-classified into overarching architectural categories[23]. We demonstrate that our method gives additional statistical certainty to conclusions regarding tissue architectures in UC from spatial omics technologies. We also demonstrate that by applying our method to CODEX we discover and localize rare cell types in UC patient biopsies. By leveraging the conditional spatial distribution of these cells within permissive, superordinate architectural structures that are more easily visualizable using scRNA-seq, we are able to conclude that these rare cells are not present in healthy control biopsies. The ability to assess both cell type supersets as well as cell architecture supersets for rare cell types provides a powerful advantage over dissociated-cell technologies. Because of the lack of explicit spatial information, although some spatial information may be implicitly encoded even in dissociated cells because of proximity-dependent changes to transcriptomes[31], dissociated-cell technologies primarily allow assessments only through cell type supersets.

Using these improved annotations, we obtain new insight into the spatial intricacies of peripheral colonic B-cell follicles in UC colonic biopsies. We identify two rare cell types in addition to five architectural components in these follicles that should be extensively verified in future CODEX experiments. While the function of the first rare cell type is unknown, the spatial restriction of these T-cells to the follicle germinal center suggests a regulatory role associated with B-cell maturation. The second rare cell type, an immature, proliferating follicle-associated B-cell population, shares markers with a subpopulation of blood-associated circulating B cells. This suggests that this fraction of circulating cells may ultimately migrate into the colon to form large, germinal center B-cell follicles, consistent with recent reports[32,33]. While this manuscript is a first but critical step towards capturing the complete cellular landscape, future functional characterization of both these cell types in large, diverse patient cohorts is crucial for identifying the role played by these cells in UC pathology. In addition to linking architectures to established patient subtypes such as therapy response, this analysis could also benefit from metabolomic profiling, spatially-resolved whole-transcriptome profiling, or spatially resolved physicochemical imaging. Although B cells play key roles in the pathophysiology of UC across peripheral, central, and mucosal immunological compartments, B-cells are relatively understudied in IBD compared to T-cells and myeloid lineage cell populations[33,34].

While B-cells have not been uniquely targeted by the approved IBD therapies, B-cells are significantly affected by current treatments and are putative targets for future IBD therapies[34]. Additional characterization of these structural components can potentially be used to develop therapies that modulate germinal center B-cell follicles in UC. We also observe differences between our healthy patient controls and our uninflamed tissues from UC patient controls, suggesting that regions designated as uninflamed may carry either residual disease-associated tissue architectures or architectures that are primed for disease flare. The formation and persistence of these architectures is potentially linked to the cycles of flare and remission commonly encountered by UC patients as reviewed in Lissner et al.[35].

Our approach should be useful for architectural interrogation, with the goal of enabling precision medicine by linking tissue architectures with disease phenotype or therapy response. While there are limitations, the majority can be mitigated through careful data collection, processing, and analysis workflows. One area that should be addressed in future implementations involves the approach used for marker thresholding. The thresholding approach assumes that marker intensities are readily thresholded; this assumption may not hold true if markers exhibit a gradient or multimodal intensity distribution. Using our CODEX dataset's CD56, for example, although a subset of B cells were clearly CD56 positive, the fact that CD56 intensities were much higher elsewhere rendered single-threshold determination ineffective for architectural feature detection (Supplemental Fig. S5).This limitation can be addressed through the use of sliding spatial windows to determine spatial patterns in overall intensity represented at the cellular level. Once the different spatial patterns are identified, each channel can be split into pseudo-channels, in which each pattern can be treated as its own marker channel for the purpose of assessing feature selection by automated methods.

Our results demonstrate the power of directly leveraging marker spatial heterogeneities for automating cell-annotation algorithms, for the detection of rare cell types, and for integrating different large datasets even when direct label transfer is not possible. As spatial omics technologies become able to assess thousands or tens of thousands of markers, manual annotation-refinement becomes formidably labor- and time-intensive[13]. With rapid advances in functional imaging, our automated approach should be able to serve as a key part of future pipelines for identifying and characterizing modular tissue architectures, as well as predicting how they locally contribute to the spectrum of health and disease.

Materials and methods

Data set acquisition

CyTOF and segmented and fluorescence-assigned CODEX datasets were acquired from the original authors[23,24]. The scRNAseq dataset was downloaded from GEO[26].

CODEX follicle extraction

Follicle ROIs were identified in FIJI based on CD19 + staining aggregation in the original images. Cell centers within each ROI were extracted in R, using the original .csv fluorescent intensity text file as well as the follicle-determined spline exported from FIJI to rapidly determine which cells were within the ROI and which were outside the ROI.

CODEX data pre-processing

Manual inspection of fluorescent image channels revealed that imaging "artifacts" were derived from two primary sources: eosinophils, which exhibited broad cytoplasmic binding to the majority of antibodies but

membrane-staining for CD15 and CD66 and smaller, punctate sources that persisted across the majority of channels. True (non-eosinophil) artifacts were identified based on broad, high-level expression of many markers except CD15 and CD66; cell center positions were retained but all marker intensities were set to 0. Due to high expression of many markers including membrane-stain signatures for CD15 and CD66, eosinophils were identified then temporarily removed from analysis. Residual artifact impacts, likely derived from segmentation, were identified by selecting a representative channel with minimal signal other than artifact, determining average signal within a non-artifact-impacted stripe, flagging all cells with signal higher than three times average background, then setting all marker intensities in flagged cells to 0.

SADIE analysis

Per fluorescent channel and using a threshold of 0.33, all cells with fluorescent signal exceeding 0.33*max were set to a value of 1 (marker-positive); all others were set to a value of 0 (marker-negative). The threshold of 0.33 was selected due to the need to remove residual signal bleed from image artifact, as well as the multimodal intensity distribution of CD45RA. SADIE analysis was then performed once for each fluorescent channel, coupled with the cell center-associated x and y coordinates, using the epiphy package implementation of SADIE and the following parameters: index = "Perry", nperm = 100, method = "shortsimplex", verbose = TRUE.

CODEX feature selection

After SADIE analysis comparing the spatial distribution of marker-positive cells to 100 random distributions using the same number of marker + cells[17] to generate a p-value (P_a), a core set of canonical cell markers was supplemented by non-extracellular matrix markers with a P_a value below 0.05 (Supplemental Table S2).

CODEX cell clustering

Initial cell clustering was performed in R, using the built-in k-means function. The initial number of clusters was determined using the fviz_nbclust implementation of elbow plots in factoextra[36]. After normalization, architecture-associated features that were marker-positive in fewer than 25 cells were given additional weight. Clusters were then displayed on the original multiplexed fluorescent image for validation using code developed by Goltsev et al.[8], with parameters tuned accordingly.

CODEX cell annotation refinement

Cell annotation refinement was performed using in-house software developed by M. Ferenc using tensorflow[37].

We built meaningful vector spaces with which we could observe well-separated clusters. Briefly, data were encoded using a one-hot approach for input into the MLP model. Principal Component Analysis, tSNE, and UMAP were used to visualize the new vector space for cluster discrimination. Data clusters were re-annotated based on the UMAP visualization.

Initial annotations for clustering were compiled into a single matrix per follicle in which each row is a single cell and the columns correspond to feature-extracted arcsinh-transformed marker fluorescent intensities, initial cluster ID, X, and Y coordinate. For initial data preprocessing, in order to address imbalances in dataset cluster sizes (Supplemental Fig. S6) 150 cells were randomly sampled from each cluster.

Model training was performed using a basic MLP (Multi Layer Perceptron) model with an input of 15 neurons corresponding to 13 fluorescent features plus the x and y coordinates, a hidden layer of 16 neurons, and an output layer of 10 neurons corresponding to the 10 desired classes for cellular annotation refinement. We performed fivefold cross validation using stochastic gradient descent with early stopping to avoid model overfitting. Cross-validation scores are shown in (Supplemental Table S4).The internal representation latent space corresponding to our 16-neuron hidden layer that emerged over the training process was then visualized using PCA, UMAP, and tSNE for dimensional reduction from 16 to 3 dimensions. Final annotation refinement was performed based on the UMAP visualization's clustering.

Generation of Voronoi images

Voronoi diagrams were created using custom code developed in Goltsev et al.[8].

Identification of CODEX Neighborhoods

For each cell in the follicle, the 10 nearest cells, including the original cell, were determined based on the annotations from CODEX cell clustering and refinement. The composition of these microenvironments was clustered using X-shift clustering with supervised annotation, using publicly available software from the Nolan lab Github. Neighborhood annotations were then re-displayed in Voronoi images.

CODEX secondary feature selection for cross-platform comparison

Features for cross-platform comparison were selected based on association with architectural sub-structures. In the event that these features were not present in the CyTOF or scRNA-seq datasets, functionally similar markers were used instead–for example, PCNA instead of Ki67. Where this was not possible, we used markers that demonstrated spatial colocalization, either directly or as a superset.

scRNAseq analysis

The R package Seurat[38] was used for analysis of scRNA-seq datasets, which had already been subjected to standard pre-processing[26]. The selection of initial parameters was guided using elbow-plots, with an initial clustering

resolution of 1.5, then further tuned based on the visualization of canonical markers on UMAP featureplots. Automated cell annotation was performed using ScType[39].

CyTOF analysis

Processed CyTOF data were obtained from the original authors and analyzed as previously described[24]. In brief, FlowJo software was utilized to gate cellular events and calculate statistics according to published conventions, and GraphPad PRISM 9 was utilized for conducting additional statistical tests and plotting figures. *P*-values were computed by unpaired Student's T-test.

Data availability

CODEX data is available at https://app.enablemedicine.com/uc-study. CyTOF data is available through contacting Dr. Aida Habtezion, corresponding author of DOI: https://doi.org/10.1038/s41467-019-10387-7. scRNA-seq data is available at the Gene Expression Omnibus under accession number GSE125527. All codes will be provide upon reasonable request to the corresponding author Dr. Stephan Rogalla.

Received: 18 February 2024; Accepted: 23 July 2024
Published online: 15 August 2024

References

1. Rutgeerts, P. *et al.* Infliximab for induction and maintenance therapy for ulcerative colitis. *N. Engl. J. Med.* **353**, 2462–2476 (2005).
2. Feagan, B. G. *et al.* Vedolizumab as induction and maintenance therapy for ulcerative colitis. *N. Engl. J. Med.* **369**, 699–710 (2013).
3. Martin, J. C. *et al.* Single-cell analysis of Crohn's disease lesions identifies a pathogenic cellular module associated with resistance to anti-TNF therapy. *Cell* **178**, 1493–1508 (2019).
4. Mitsialis, V. *et al.* Single-cell analyses of colon and blood reveal distinct immune cell signatures of ulcerative colitis and Crohn's disease. *Gastroenterology* **159**, 591–608 (2020).
5. Schürch, C. M. *et al.* Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. *Cell* **182**(5), 1341–1359 (2020).
6. Ali, H. R. *et al.* Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer. *Nat. Cancer* **1**(2), 163–175 (2020).
7. Kuswanto, W., Nolan, G. & Lu, G. Highly multiplexed spatial profiling with CODEX: Bioinformatic analysis and application in human disease. *Semin. Immunopathol.* **45**, 145–157 (2023).
8. Goltsev, Y. *et al.* Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. *Cell* **174**, 968–981 (2018).
9. Pasquini, G., Arias, J. E. R., Schäfer, P. & Busskamp, V. Automated methods for cell type annotation on scRNA-seq data. *Computat. Struct. Biotechnol. J.* **19**, 961–969 (2021).
10. Brbić, M. *et al.* Annotation of spatially resolved single-cell data with STELLAR. *Nat. Methods* **19**, 1411–1418 (2022).
11. Adler, M., Kohanim, Y. K., Tendler, A., Mayo, A. & Alon, U. Continuum of gene-expression profiles provides spatial division of labor within a differentiated cell type. *Cell Syst.* **8**, 43–52 (2019).
12. Newman, E. A., Kennedy, M. C., Falk, D. A. & McKenzie, D. Scaling and complexity in landscape ecology. *Front. Ecol. Evol.* **7**, 293 (2019).
13. Zhang, W. *et al.* Identification of cell types in multiplexed in situ images by combining protein expression and spatial information using CELESTA. *Nat. Methods* **19**, 759–769 (2022).
14. Wu, J. & Loucks, O. L. From balance of nature to hierarchical patch dynamics: A paradigm shift in ecology. *Q. Rev. Biol.* **70**, 439–466 (1995).
15. Holling, C. S. Cross-scale morphology, geometry, and dynamics of ecosystems. *Ecol. Monogr.* **62**, 447–502 (1992).
16. Wolpert, D., Libby, E., Grochow, J. A. & Dedeo, S. The many faces of state space compression. In *From Matter to Life: Information and Causality* (eds Walker, S. *et al.*) 199–243 (Cambridge University Press, 2017).
17. Perry, J. N. Spatial analysis by distance indices. *J. Anim. Ecol.* **64**, 303–314 (1995).
18. Perry, J. N., Winder, L., Holland, J. M. & Alston, R. D. Red–blue plots for detecting clusters in count data. *Ecol. Lett.* **2**, 106–113 (1999).
19. Li, B., Madden, L. V. & Xu, X. Spatial analysis by distance indices: An alternative local clustering index for studying spatial patterns. *Methods Ecol. Evol.* **3**(2), 368–377 (2012).
20. Zhao, X. & van Praag, H. Steps towards standardized quantification of adult neurogenesis. *Nat. Commun.* **11**, 4275 (2020).
21. Waise, S. *et al.* An optimised tissue disaggregation and data processing pipeline for characterising fibroblast phenotypes using single-cell RNA sequencing. *Sci. Rep.* **9**, 9580 (2019).
22. Soteriou, D. *et al.* Rapid single-cell physical phenotyping of mechanically dissociated tissue biopsies. *Nat. Biomed. Eng.* **7**, 1392–1403 (2023).
23. Mayer, A. T. *et al.* A tissue atlas of ulcerative colitis revealing evidence of sex-dependent differences in disease-driving inflammatory cell types and resistance to TNF inhibitor therapy. *Sci. Adv.* **9**, eadd1166 (2023).
24. Rubin, S. J. S. *et al.* Mass cytometry reveals systemic and local immune signatures that distinguish inflammatory bowel diseases. *Nat. Commun.* **10**(1), 2686 (2019).
25. Kharchenko, P. V. The triumphs and limitations of computational methods for scRNA-seq. *Nat. Methods* **18**, 723–732 (2021).
26. Boland, B. S. *et al.* Heterogeneity and clonal relationships of adaptive immune cells in ulcerative colitis revealed by single-cell analyses. *Sci. Immunol.* **5**, eabb4432 (2020).
27. Turner, M. G., Gardner, R. H. & O'Neill, R. V. In *Landscape Ecology in Theory and Practice* (ed. Golley, F. B.) 1–7 (Springer, 2001).
28. Slovin, S. *et al.* Single-cell RNA sequencing analysis: A step-by-step overview. In *RNA Bioinformatics. Methods in Molecular Biology* Vol. 2284 (ed. Picardi, E.) 343–365 (Humana, 2021).
29. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* **42**, 293–304 (2023).
30. Garrido-Trigo, A. *et al.* Macrophage and neutrophil heterogeneity at single-cell spatial resolution in human inflammatory bowel disease. *Nat. Commun.* **14**, 1 (2023).
31. Cang, Z. & Nie, Q. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nat. Commun.* **11**, 2084 (2020).
32. Pararasa, C. *et al.* Reduced CD27− IgD− B cells in blood and raised CD27− IgD− B cells in gut-associated lymphoid tissue in inflammatory bowel disease. *Front. Immunol.* **10**, 361 (2019).
33. Uzzan, M. *et al.* Ulcerative colitis is characterized by a plasmablast-skewed humoral response associated with disease activity. *Nat. Med.* **28**, 766–779 (2022).
34. Castro-Dopico, T., Colombel, J. F. & Mehandru, S. Targeting B cells for inflammatory bowel disease treatment: Back to the future. *Curr. Opin. Pharmacol.* **55**, 90–98 (2020).

35. Lissner, D. & Siegmund, B. Ulcerative colitis: Current and future treatment strategies. *Dig. Dis.* **31**, 91–94 (2013).
36. Kassambara, A. and Mundt, F. Package 'factoextra': Extract and visualize the results of multivariate data analyses. (2016).
37. Abadi, M. et al. TensorFlow: A system for large-scale machine learning. *12th USENIX symposium on operating systems design and implementation*, OSDI16, 265–384 (2016).
38. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
39. Ianevski, A. *et al.* Fully-automated and ultra-fast cell-type identification using specific marker combinations from single cell transcriptomic data. *Nat. Commun.* **13**, 1 (2022).

Acknowledgements

We thank T.L and H.Y.H for many intriguing conversations and guidance.

Author contributions

D.R.H., S.J.SR, M. F, E.A.H, and S.R conceptualized the manuscript. D.R.H., S.J.SR, M. F, E.A.H, B.S.B., G.P.N. and S.R drafted the manuscript. D.R.H, S.J.SR. M.F. E.A.H, J. T. C, R.D, S.R. and A.N.K were involved in analysis design and data analysis. All authors critically reviewed and edited the final version of the manuscript.

Funding

Research reported in this publication was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under award number UL1TR003142 and by the Kenneth Rainin Foundation, 2018-575(S.R.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-68397-5.

Correspondence and requests for materials should be addressed to D.R.H. or S.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.