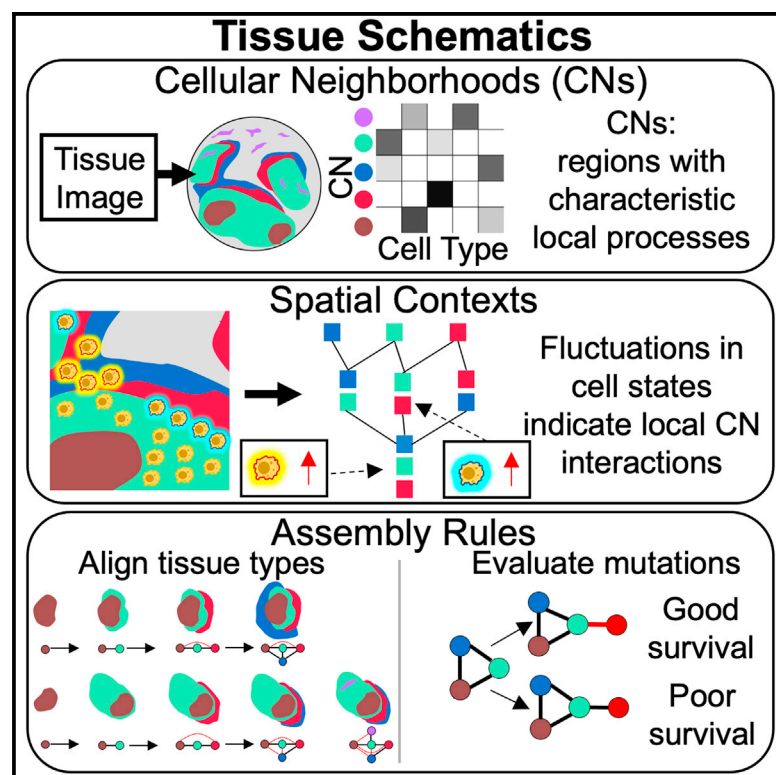# Tissue schematics map the specialization of immune tissue motifs and their appropriation by tumors

Graphical abstract

Authors

Salil S. Bhate, Graham L. Barlow, Christian M. Schürch, Garry P. Nolan

Correspondence

gnolan@stanford.edu

In brief

We present Tissue Schematics, a computational and conceptual framework for using high-parameter imaging data to map how tissues are formed by the assembly of cellular neighborhoods into tissue motifs. Applying our framework to human lymphoid tissues and the colorectal cancer immune-tumor microenvironment suggests mechanistic insights into the specific biology of these tissues, as well as general principles of tissue organization.

Highlights

- Computational framework for reverse-engineering tissues from high-parameter imaging data

- Tissue schematics map modular assembly of cellular neighborhoods into tissue motifs

- Aligning human lymphoid tissues reveals core assembly specialized by each tissue type

- Mutations of a higher-order motif associated with patient survival in colorectal cancer

CellPress

Cell Systems

Article

Tissue schematics map the specialization of immune tissue motifs and their appropriation by tumors

Salil S. Bhate,[1,2,3,6] Graham L. Barlow,[1,2,4,6] Christian M. Schürch,[1,2,5] and Garry P. Nolan[1,2,7,*]
[1]Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford University, Stanford, CA 94305, USA
[2]Department of Pathology, Stanford University School of Medicine, Stanford University, Stanford, CA 94305, USA
[3]Department of Bioengineering, Stanford University Schools of Medicine and Engineering, Stanford University, Stanford, CA 94305, USA
[4]Program in Immunology, Stanford University School of Medicine, Stanford University, Stanford, CA 94305, USA
[5]Department of Pathology and Neuropathology, University Hospital and Comprehensive Cancer Center Tübingen, Tübingen, Germany
[6]These authors contributed equally
[7]Lead contact
*Correspondence: gnolan@stanford.edu
https://doi.org/10.1016/j.cels.2021.09.012

SUMMARY

A schematic of a biological system, i.e., a representation of its pieces, how they are combined, and what they do, would facilitate understanding its essential organization and alteration in pathogenesis or evolution. We present a computational approach for constructing tissue schematics (TSs) from high-parameter imaging data and a biological model for interpreting them. TSs map the spatial assembly of cellular neighborhoods into tissue motifs, whose modular composition, we propose, enables the generation of complex outputs. We developed our approach in human lymphoid tissue (HLT), identifying the follicular outer zone as a potential relay between neighboring zones and a core lymphoid assembly with modifications characteristic of each HLT type. Applying the TS approach to the tumor microenvironment in human colorectal cancer identified a higher-order motif, whose mutated assembly was negatively associated with patient survival. TSs may therefore elucidate how immune architectures can be specialized and become vulnerable to reprogramming by tumors.

INTRODUCTION

It is a tenet of evolution that prior evolutionary success can be repurposed for variations on function or for the creation of new functions. Thus, one expects there to be repeating units of architecture serving similar purposes in multiple tissue types and disease contexts. However, it is less clear the extent to which such units combine to generate more complex ones and how this assembly contributes to tissue biology and pathology.

We present here "tissue schematics" (TSs), a conceptual and algorithmic framework for defining and quantifying a tissue's architectural units and assembly from high-parameter imaging data (Angelo et al., 2014; Ståhl et al., 2016; Agasti et al., 2017; Wei et al., 2017; Goltsev et al., 2018; Lin et al., 2018; Saka et al., 2019; Vickovic et al., 2019). In our framework, a TS of an organ consists of three components (which are detailed and biologically interpreted in the results section): (1) the collection of cellular neighborhoods (CNs) (each defined by a characteristic local composition of cell types), (2) a map of distinct microenvironments formed by CNs that are co-localized (termed "spatial contexts" [SCs]), and (3) a map of hierarchically formed collections of CN regions with specific SCs between them (which we

term "motifs") and the rules governing their formation (termed assembly rules [ARs]). We have compiled a glossary of the concepts utilized herein (Table 1).

One uniquely diverse family of tissues with a common general purpose is human lymphoid tissue (HLT) which includes tonsils, lymph nodes, spleens, and others. These are tissues wherein the immune system carries out critical behaviors that depend on co-ordinated interactions occurring between specialized cell types. While different types of lymphoid tissue share some common functions (e.g., antigen presentation and antibody production), they each are specialized for certain tasks (e.g., the spleen is better suited to respond to blood-borne pathogens than lymph nodes). This family of tissues therefore provides an ideal system for exploring how units of architecture are utilized in multiple contexts and how complex immune functions are achieved by composing them to generate more complex functional units at the tissue level.

We developed and validated the TS concepts and algorithms on a CODEX dataset of two tonsils, a spleen, and a lymph node (Kennedy-Darling et al., 2021). First, we identified a collection of CNs for HLT and validated their suitability for constructing TS, in accordance with the TS's underlying biological model. Next, we

Table 1. Glossary of terms

Term	Abbreviation	Definition
Cell type	CT	A CT is a set of cells in the imaging data defined by clustering cells with respect to their expression of phenotypic markers.
Window of size k		The window of a cell is the set of its k-nearest neighbors (in terms of spatial distance) in the imaging data.
Cellular neighborhood	CN, cnX	A CN is a region of the tissue with a homogeneous local composition of CTs and approximated by clustering cells with respect to the cell type composition of their windows.
Spatial context	SC(cnX,cnY,cnY,…)	A spatial context is a region of the tissue where the local processes of different CNs could be interacting. We approximate them as CN-combinations, regions where cells assigned to multiple CNs are in close proximity. These are labeled by a set of CNs. A cell is assigned to a SC for a set of CNs, say $CN_1,…, CN_n$ if more than 90% of the cells in its window of size 100 are assigned to one of those n CNs and if it is the minimal set of CNs that has this property (i.e., removing any CN_i from the combination, the total number of cells in its window assigned to any of the CNs except from CN_i is less than 90%).
CN instance		A CN, being a region of the tissue, need not be connected as a spatial domain. We refer to its connected components as its instances. These are identified by constructing the 10-nearest neighbor graph between cells and identifying its connected components in the CRC dataset or using image-processing libraries for identifying connected components of Boolean images in the HLT dataset.
Tissue graph		This is a colored graph formed for each tissue, in which the vertices represent instances of CNs, vertex colors represent the CN, and edges represent spatial proximity (defined by having neighboring cells).
Motif	A, B, C, …	A motif is a (typically small) colored graph whose vertices are labeled by CN names and is the basic repeated structure we search for in the tissue graph. We have only considered motifs in which a single CN can appear only once.
Two-chain	cnX-cnY	a motif consisting of two CNs with one edge.
Triangle	Triangle(cnX,cnY,cnZ)	a motif consisting of three CNs all connected.
Motif instance		A motif instance is a subgraph of a tissue graph isomorphic to a motif (i.e., a collection of CN instances and adjacency relationships that agree with the edges of the motif).
Submotif		Viewing a motif as a colored graph, a colored subgraph of it is referred to as a submotif.
Extension of a motif instance		Given a motif A and a submotif B, we say that an instance of A extends to an instance of B if the instance of A is a colored subgraph of the instance of B in the tissue graph. Note that an instance of a given motif could extend to an instance of another in multiple ways.
Assembly rule	AR, A → B	An AR specifies that all (approximated here by a high proportion of) the instances of a motif A extend to instances of another motif B where A is a submotif of B.
Basic rule		Given a motif B, and a submotif A, if there is an AR A→ C, meaning that instances of the motif A can be extended to instances of the motif C, we can infer that an instance of B can always be extended to the motif formed by adding on the piece to A that extends it to C. Basic rules are ARs that cannot be inferred from simpler rules in this way.
Rule graph		Given the collection of ARs, we construct a graph that represents these rules, wherein paths in the graph correspond to composite inferences.
Maximum entropy null distribution		Given a tissue graph, we consider the level 0 null set of its CN assignments obtained by arbitrary transpositions of vertex colors, and the level 1 null set obtained by transpositions between vertices whose neighboring vertices have the same CN assignments. The maximum entropy null distribution is the uniform distribution on these null sets of colorings.

(Continued on next page)

Cell Systems
Article

CellPress

developed an algorithm to identify SCs—tissue regions, in which local processes of different CNs could be interacting and a visualization termed a "CN combination map" (CNM). The CNM of HLT revealed structures consistent with the known function of immune architecture as well as suggested that interactions between local processes of CNs in SCs could be identified by phenotypic changes in cell types. In the biological model underlying TS, motifs (collections of CN regions with specific SCs between them), could have emergent functionality arising from signal propagation between CNs. We therefore identified the motifs in HLT and the rules governing their assembly in each tissue type. We used these rules to align the architectures of the tonsil, lymph node, and spleens and to quantify their complexity.

Our analyses of HLT revealed a core lymphoid motif and its specialization by each tissue type. We also found that the tonsillar architecture was most complicated, followed by lymph node and then spleen. In addition, the assembly of the tonsil and lymph node was more similar than either's was to the spleen.

We had yet to establish whether the SCs and motifs as we had defined could have a functional role, which we could infer if specific SCs and motifs were statistically associated with disease outcome. Therefore, we applied the TS analysis approach to a previously reported CODEX dataset of the immune-tumor microenvironment (iTME) of human colorectal cancer (CRC) (Schürch et al., 2020). In this dataset, which captures two extremes of the CRC iTME spectrum, patients can be stratified by two characteristic types of immune infiltrate in their iTMEs. One patient group, termed Crohn's-like reaction (CLR) have iTMEs in which tertiary lymphoid structures (TLSs, resembling the follicles of lymphoid organs) form and who have significantly better survival outcomes than patients of the diffuse inflammatory infiltration (DII) patient group, whose iTMEs do not have TLS.

The first ingredient of a TS, identification of CNs, had been previously performed (Schürch et al., 2020). Next, we not only mapped the SCs formed by these CNs (and thus, which local processes could be interacting) as we had in the HLT dataset, but also assessed what the result of these local interactions could be by identifying which SCs were associated with changes in expression levels of functional markers on key immune subsets. Depicting these results on the CNM for the CRC iTME revealed a rich map highlighting potential functional roles of its CNs and their local interactions. For example, it suggested that, similar to HLT, the TLS could likely be involved in antigen-dependent, helper-T-cell-mediated proliferation of B cells and a potential interaction between the characteristic local processes of the tumor and tumor boundary that leads macrophage to proliferate.

We next turned to the assembly of motifs in the CRC iTME. We mapped its CNs to those shared with HLT, finding conserved assembly rules there. We developed a statistical way to quantify motifs that were "higher-order," i.e., actively assembled by biological programs counteracting entropy. This revealed a higher-order tissue motif of 3 CNs—consisting of contiguously assembled regions of the T cell enriched, macrophage enriched, and vasculature neighborhoods assembled, into which the tumor CN was inserted, exclusively in the high-risk DII patient group. The insertion of the tumor into this motif was associated with worsened survival outcomes within the high-risk patient group. Crucially, its submotifs were not associated with survival. Thus, our results confirm that the TS approach provides a way to identify and dissect emergent tissue functions from modular construction of complex tissue structures from simpler interacting, multicellular components to gain potential mechanistic insights into antitumoral immunity.

Ours is not the only possible notion of a tissue's "schematic" that might be applied. Multiple notions of tissue component and corresponding notion of interaction have yielded insights into tissue function from high-parameter imaging data (such as cells Schapiro et al., 2017; Goltsev et al., 2018; Keren et al., 2018; Arnol et al., 2019, i-niches Goltsev et al., 2018, cellular neighborhoods [CNs] Schürch et al., 2020; Stoltzfus et al., 2020, and communities Jackson et al., 2020). However, the primary findings as outlined above and herein were enabled by and hence justify our TS formalization. These results highlight the utility of quantitatively exploring compositional tissue assembly with TS for both basic and translational insights into tissue function and malfunction.

RESULTS

CNs as primitive components for tissue schematics

In our framework, a TS is a computational description of how a tissue is built from its components. Abstractly, any schematic of a tissue can be thought of as a description of a tissue with respect to a model for (1) what the components of a tissue can be, (2) how more complex components are formed from simpler components, and accordingly, (3) how these more complex components achieve biological functions through interactions between those of their subcomponents. The components we focus on range in complexity; they include cell types and cell structures such as vessels, but also include more complex objects such as follicles and follicle super-structures. We focus on spatial contact as the means by which the components form more complex components. The details of the components, as well as the method by which tissue components achieve complex functions through interactions between those of their subcomponents are depicted in Figure 1A and detailed presently.

To start, we assumed that tissues would be comprised of distinct types of regions, each defined by ongoing characteristic local, molecular processes (Figure 1A, left column). This means that in the vicinity of a given point in such a region, characteristic biological signals can be generated, sent, received, and then processed to an outcome (for example, cells might differentiate in response to a ligand, or molecules might be secreted). As
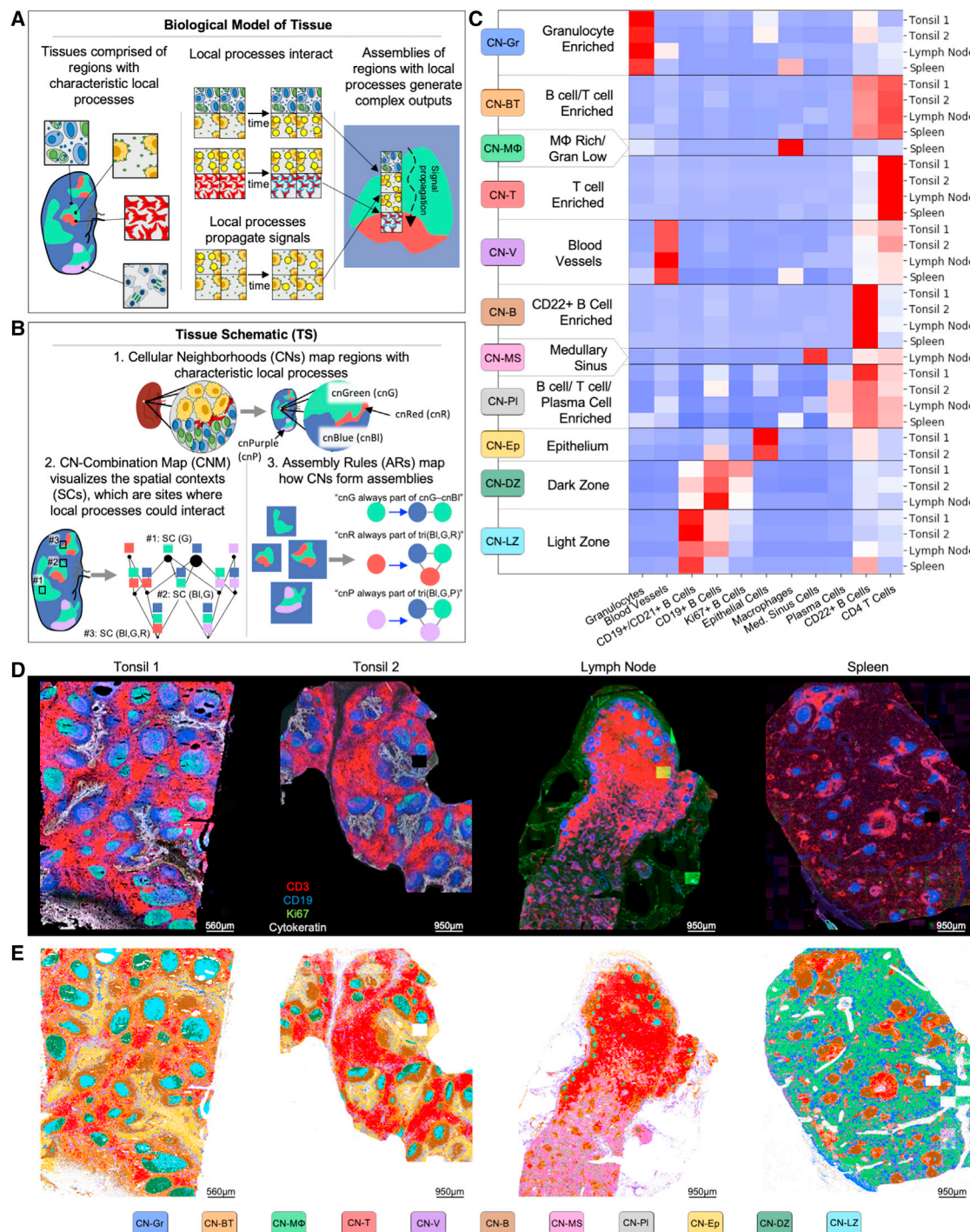
Figure 1. Cellular neighborhoods are building blocks for tissue schematics

(A) Cell neighborhood model of tissue behavior underlying TSs. Left column: tissues comprised of different types of regions, each with characteristic local processes ongoing. Middle column, top: these local processes of CNs can interact, giving rise to new signals where they do. Middle column, bottom: local processes can propagate signals, for example by cellular movement followed by secretion, direct diffusion or cell-cell signaling. Right column: assemblies of connected tissue regions of different types can generate complex outputs by combining signal propagation by and interactions between the characteristic local processes of each region (meaning that a given such region can act as a relay between others).

(B) Computational representation utilized by tissue schematics to map cellular architecture in accordance with model in (A). Top row: cell neighborhoods are identified. Bottom row, left: the regions where CNs are in contact are termed "spatial contexts" and represented as a CN combination map (CNM). Bottom row, right: assembly rules (ARs) describing how collections of CN regions form assemblies are mapped.

(legend continued on next page)

Cell Systems
Article

such, at given points in a tissue, in whose vicinity there are multiple local processes ongoing, these processes could interact to generate additional distinct biological signals (Figure 1A, middle column, top row). Such signals could be propagated, for example, by direct molecular diffusion, cellular movement followed by secretion, or cell-cell signaling (Figure 1A, middle column, bottom row). Consequentially, an "assembly" of regions—multiple connected regions in spatial proximity—would generate complex outputs by coupling signal propagation and processing by local processes, as well as the interactions between local processes around points where regions are in close proximity (Figure 1A, right column). For example, the green region in (Figure 1A, right column), acts as a relay, propagating the yellow signal through it from the interface with the blue region (where it is generated by interactions between the local processes of the green region and blue region) to the red region (where there is a local interaction corresponding to the yellow signal triggering activation of the red cells). Thus, it is the premise here that by identifying such regions (and potential signals generated when they are in local proximity), as well as identifying how related assemblies are associated with distinct outcomes, we would obtain a map of the different component regions and derive how they might interact to generate complex functions. How then can a tissue be computationally described with respect to the outlined biological model for a tissue schematic?

Since the local processes in our biological model would likely be mediated by cells and cell-cell interactions, we approximated the regions with such processes as those regions that had a characteristic local composition of cell types, termed CNs (Figure 1B, top row) as previously defined and applied (Schürch et al., 2020). In the pseudo-tissue of the diagram, there are four CNs: cnGreen, cnRed, cnBlue, and cnPurple. As noted above, we termed regions where there are local spatial contacts between CNs as "spatial contexts" (SCs), since these are likely to be the sites where the characteristic local processes of each CN interact. As such, the nature of these interactions can be inferred from the phenotypes of cells within these regions. In a TS, we describe the collection of SCs in a tissue with a CNM (Figure 1B, bottom row, left column). The borders of CNs are inherently challenging to define. While we had to delineate CNs with discrete boundaries for computational purposes, describing the tissue with respect to SCs enables explicitly defining the dynamic regions in which such interactions could occur, thereby providing one approach to solve this problem.

In accordance with the model above, it would be expected that a collection of connected CN regions, with specific SCs forming between them (which we term an "assembly"), would be poised to generate unique biological signals by coupling the ability of the local processes of each CN to interact and propagate signals. Assuming that evolution repurposes common elements to achieve diverse function, a given CN (or sub-assembly) could be expected to be used in multiple distinct assemblies. Thus, we would expect there to be an underlying order to how

the CNs form assemblies, with variations of this order utilized, or altered, by functionally similar tissues in different contexts (such as in different lymph nodes or various pathologies). Accordingly, in a TS, we identify "motifs," which are certain assemblies present in multiple locations in the tissue and a collection of "assembly rules" (ARs) capturing how motifs are assembled (Figure 1B, bottom row, right column). Thus, a TS describes how a tissue is built, by identifying CNs, the SCs they form, and the rules governing their assembly. While all the concepts we introduce will be fully explained and developed below, we have compiled in Table 1, for conciseness and reference, a glossary of all the terms and their abbreviations used.

TSs of HLT were constructed from 46-parameter images of two tonsils, a lymph node, and a spleen, with data collected using the imaging technology co-detection by indEXing (CODEX; Kennedy-Darling et al., 2021). These images captured large areas of the tissue samples (tonsil #1: 32 mm^2, tonsil #2: 34 mm^2, lymph node: 26 mm^2, spleen: 49 mm^2), providing a comprehensive view of the architecture of HLT.

Using approaches as previously described (Goltsev et al., 2018; Schürch et al., 2020), 25 cell types were identified in HLT following single-cell segmentation and clustering according to marker expression (Figure S1A, cell types in heatmap). Next, CNs were identified as previously reported (Schürch et al., 2020; Stoltzfus et al., 2020). For each cell, a "window" of size 20 was extracted consisting of a cell and its 19 nearest neighbor cells, defined by Euclidean distance between cell centroids in X, Y space (a window size of 20 was chosen since a trial of the process below with a window size of 10 or 30 did not substantially alter the observed CNs [Figure S2], confirming that the CNs were persistent across small changes in length scale). Windows were next clustered with respect to the frequencies of these 25 cell types. The clusters resulting were validated as follows. First, to determine whether there was substructure within these CNs that had not been identified by the clustering algorithm, we examined the distribution of cell-type density per window in the windows assigned to each CN within each tissue (Figure S1B).

None of the CNs exhibited multi-modal distributions, demonstrating that each cell type was relatively evenly distributed throughout the region assigned to a given CN. Second, the clusters were manually inspected by overlaying the cells allocated to them over select channels from the raw tissue images, for each tissue individually, to ensure that they captured distinguishable tissue regions defined by their corresponding local composition of cell types (Figure S3), and clusters were merged accordingly. This resulted in 11 clusters with distinct cell-type compositions comprising the collection of CNs in HLT (Figures 1C and S1A). Each CN was named with a one or two letter acronym that represented the characteristic cellular composition of that CN and was assigned a color that remains consistent throughout the figures. The cellular compositions per tissue for each of the 11 CNs are displayed alongside their color assignment and name. For

(C) Heatmap of average window compositions of CNs in each tissue sample, normalized across columns. Red indicates enrichment of a cell type in windows assigned to a CN in a given tissue. Heatmaps are shown only for tissues in which a given CN was observed.

(D) Images of each tissue sample showing T cells (CD3, red), B cells (CD19, blue), epithelial cells (cytokeratin, white), and proliferative cells (Ki67, green). Scale bars indicated for each tissue.

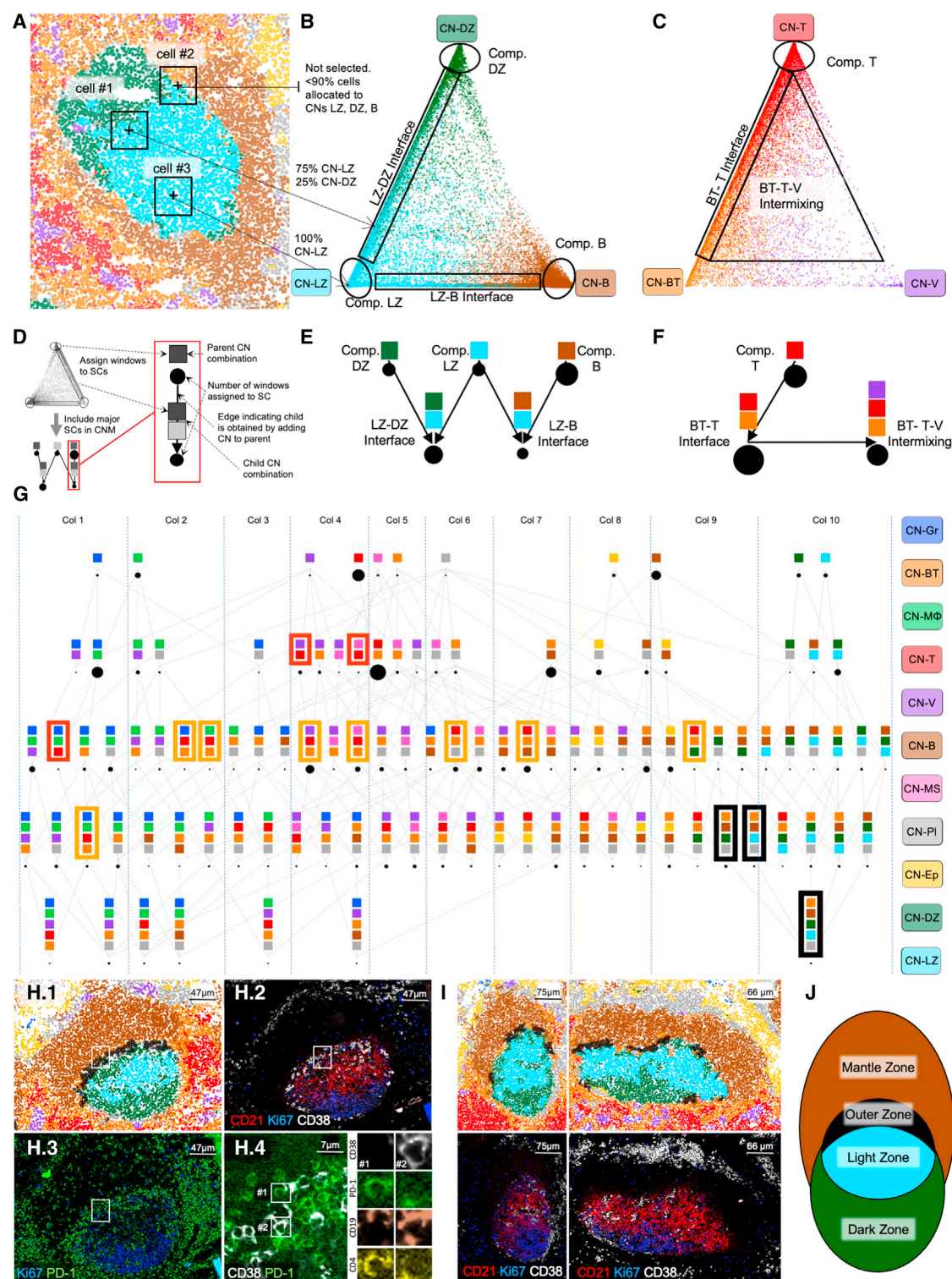(E) Cells in each tissue colored by CN assignment.

Figure 2. Spatial contexts are sites of functional interactions between local processes of CNs

(A) Cells in a follicle from tonsil 1 colored by CN assignment (see legend to the right of G). Cells (indicated by +) and their windows (boxed regions) highlight different SCs.

(B) Barycentric coordinate projection of windows with greater than 90% of cells assigned to CNs B, LZ, or DZ. Outlined regions on edges and vertices indicate abundant SCs (compartmentalized regions of CN and interfaces).

(C) Barycentric coordinate projection of windows with greater than 90% of cells assigned to CNs BT, T, or V (compartmentalized regions of CN and interfaces).

clarity, cell types that were not enriched in any CN were omitted from the heatmap in Figure 1C, but the full heatmap is shown in Figure S1A. We acknowledge and expect that their presence could have meaning in a yet to be defined context beyond the scope of this report. In Figure 1C, each CN is broken into rows corresponding to cellular composition in each tissue, showing that each CN has a similar cellular composition in all HLT types.

The 11 CNs in HLT are described here and shown consecutively in Figure 1C by row. Seven of the CNs are found in all tissues and consist of the following: (1) cnGr, which is enriched in granulocytes; (2) cnBT, which is enriched in primarily B cells and CD4$^+$ T cells; (3) cnT, which is strongly enriched in CD4$^+$ T cells; (4) cnV, which is enriched in blood vessels; (5) cnB, which is enriched in CD22$^+$ B cells; (6) cnPl, which is similar to cnBT in that it is enriched in both B cells and CD4$^+$ T cells but distinct in that it is also enriched in plasma cells and blood vessels; and (7) cnLZ, which is enriched in CD19$^+$/CD21$^+$ B cells and is recognizable as the follicular light zone. Four CNs were only found in a subset of the HLT samples. cnMS was found only in the lymph node and corresponds to the medullary sinus. cnMΦ was only found in spleen and is enriched in macrophages. Note that cnGr in the spleen also involves macrophages, thus cnMΦ is distinct in the spleen from cnGr in part due to its lack of granulocytes. cnEp contains epithelial cells and was found only in the tonsils. cnDZ contains a high frequency of CD19$^+$ B cells and Ki67$^+$ (proliferating) B cells and is recognizable as the follicular dark zone. cnDZ is found in the lymph node and tonsils but not in the spleen sample that was imaged.

In the raw tissue images, the B cell follicles, germinal centers, and T cell zones are visible in each tissue from the CD19, Ki67, and CD3 staining, respectively (Figure 1D). Visualization of the CNs by plotting cell centers as points colored by their CN assignment demonstrated that the CNs capture and represent tissue regions recognized in classical hematopathology (Figure 1E). Seven out of eleven CNs were found in all four tissues whereas cnEp, cnMS, and cnMΦ were specific to tonsil, lymph node, and spleen, respectively. Therefore, these CNs are a suitable approximation to regions of HLT with characteristic local processes that we can use to initiate the construction of TSs of HLT. How are CNs composed to generate the complex functions of HLT?

Spatial contexts as sites of possible interactions between local tissue processes

We reasoned that the regions of the tissue where CNs are in close contact would be the locations where the characteristic local processes of each CN could interact through local molecular reactions. Such interactions might be reflected in specialized biological processes arising exclusively in such regions. For example, interfaces between CNs are sites where a molecule from one CN might contact cells from another CN that enables unique proliferation events or other cell activation processes (Figure 1A, middle column). As per above, we termed such regions SCs. Assuming that the signals generated in individual CNs played a role in complex tissue outputs, the SCs that are present, as well as the specialized local processes occurring within them, could act as the "relays" linking together the individual components in our tissue schematic. How, therefore, might we systematically identify and represent the SCs in each tissue?

We began by investigating the SCs formed by the CNs of B cell follicles, which are dense regions of B cells and supporting cells that are critical for B cell activation and antibody production. In many ways, such follicle-like structures are a signature component of lymph node function and structure. Follicular structures are found in most lymphoid organs and arise in peripheral sites such as inflamed tumors (Binnewies et al., 2018) and sites of autoimmunity (Pipi et al., 2018). Accordingly, all tissues in our dataset contained follicles with characteristic CD19 staining (Figure 1D). The main CNs of the follicles are cnB, cnDZ, and cnLZ as illustrated by a representative follicle (Figure 2A). Some cells in cnLZ are located at an interface of cnLZ and cnDZ (Figure 2A, e.g., cell #1), other cells in cnLZ contact several other CNs (Figure 2A, e.g., cell #2 contacting cells in cnLZ, cnB, cnBT, cnDZ, and cnPl), and some are surrounded by only cnLZ cells (Figure 2A, e.g., cell #3). This example illustrates that a given collection of CNs does not form all possible SCs that could be but instead forms a characteristic collection reflective of the requirements of the tissue. Traditionally, the borders of CNs have been inherently hard to define but these initial analyses suggested that these regions could be biologically significant. Therefore, the question of to which CN the cells of a given SC should be allocated could be subverted by assessing the unique processes occurring therein.

The composition of a cell's nearest neighbors (with respect to their CN assignments) should reflect the SC within which it is located. We used this to identify the SCs across all follicles in our dataset. Specifically, windows containing the 100 nearest spatial neighbor cells around a given index cell were extracted. This window size—corresponding to a radius of approximately 5 cells—was used as a conservative length scale on which biologically relevant cues could be shared between CNs (based

(D) Legend of components of CNM. Size of black circle indicates abundance of combination. Colors of squares indicate the CNs in combination. Edges indicate the child combination is obtained from the parent combination by adding one CN.
(E) CNM of follicle region.
(F) CNM of T cell zone region.
(G) CNM of combinations comprising greater than 0.001% of total cells. Red rectangles indicate SC containing cnT but not cnBT. Orange rectangles indicate SC that contain co-localized combinations of cnT and cnBT with a third CN. (cnBT, cnT, cnMΦ, and cnGr) is included in this set to facilitate comparing its abundance with that of (cnT, cnMΦ, and cnGr). Columns are overlaid to facilitate reference in the main text.
(H) Tissue images of follicle to illustrate SCs from CNM. (H.1) CD21, Ki67, and CD38 expression in follicle region. (H2) Ki67 and PD-1 in region. (H.3) Cells colored by CN assignment in same region as shown in images H.1 and H.2. Black points indicate cells assigned to the combinations (cnBT, cnPl, cnDZ, and cnB), (cnBT, cnPl, cnLZ, and cnB), or (cnBT, cnPl, cnLZ, cnDZ, and cnB). These combinations are indicated with black boxes in (G). (H.4) CD38 and PD-1 are expressed on different cell types within region outlined in white box in H.1 and H.2. Co-staining of two cells outlined in white boxes demonstrate CD38$^+$ staining is restricted to CD19$^+$ B cells and PD-1 is restricted to CD4$^+$ T cells.
(I) CN plots of two other follicles from the same patient showing that the same contexts (bottom row) do not have enriched CD38 expression within them as seen in corresponding tissue image (top row). (J) Schematic illustrating location of outer zone in follicle.

on, for example, that intercellular IL-2 signaling occurs between immune cells situated at distances ranging from 1 to 10 cell diameters apart; Oyler-Yaniv et al., 2017). Next, cells whose windows were comprised of predominantly (>90%) of cells allocated to cnB, cnDZ, or cnLZ were selected (excluding, for example, cell #2 from Figure 2A). Finally, the CN compositions of the windows of these cells were projected into the barycentric coordinate system (Figure 2B, the window of each cell represented by a point in the triangle).

The distribution of cells in this coordinate system highlights which SCs are formed by the CNs that are indicated at each vertex. A cell is close to a vertex if its window is comprised of predominantly cells assigned to the CN indicated by that vertex (Figure 2A, e.g., cell #3 surrounded only by cells assigned to cnLZ). Therefore, increased density around a vertex indicates the presence of an SC that corresponds to a population of cells "compartmentalized" within that CN—i.e., those which are surrounded by cells only of the same CN. Accordingly, a cell is close to an edge of the triangle if its window is predominantly comprised of cells allocated to the two CNs connected by that edge (Figure 2A, e.g., cell #1 with a window containing 75% cells assigned to cnLZ and 25% cells assigned to cnDZ). Thus, the density of points along an edge indicates the presence of an SC corresponding to a population of cells at the interface between those two CNs. Cells in the center of the triangle have windows comprised of cells assigned to all three CNs.

The SCs of the follicles consisted of three compartmentalized regions (black circles around vertices labeled "Comp. LZ," "Comp. DZ," and "Comp. B") and two interface regions (black rectangles labeled "LZ-DZ Interface" and "LZ-B Interface") (Figure 2B). Comparison of the observed SCs to the established behavior of the GC during antibody affinity maturation highlights the important biological information present within the SCs formed by a collection of CNs. First, CNs LZ and DZ being compartmentalized reflects how the tissue segregates proliferating B cells in the dark zone and T cells in the light zone to increase the likelihood of finding a higher affinity clone by allowing B cell clones to expand briefly without selective pressure, while maintaining the stringent selection in the light zone (Kepler and Perelson, 1993). Second, a large SC with these two CNs reflects how the tissue efficiently cycles B cells between the selection and expansion stages in the light zone and dark zone. The fact that the SC formed by cnDZ and cnB was small could interpreted either that their direct contact is less critical to HLT function or that their active separation is critical to the outputs of HLT.

Another important region in HLT, the T cell zone, where antigen-presenting cells prime T and B cell responses, was similarly analyzed for its SCs. The T cell zone is well described by cnBT, cnT, and cnV. We selected cells, as we had with the follicular region, whose windows were comprised predominantly (>90%) of cells assigned to cnBT, cnT, or cnV to describe the T cell zone. Here, the cells were sparse near the vertices corresponding to CNs V and B—but were dense near the vertex corresponding to cnT (Figure 2C). Furthermore, the only edge that had density along it was that between CNs BT and T (Figure 2C, black rectangle labeled "BT-T Interface"). A higher density of cells was distributed within the volume (off the vertices and edges) of the T cell zone's triangle than within the area of the follicles' triangle—indicating substantial co-localization of all three cnT, cnBT, and cnV (Figure 2C, black triangle labeled "BT-T-V Intermixing"). The SCs found within the follicle and T cell zone capture two types of local interaction structure possible with three CNs: the diffuse nature of cnT, cnV, and cnBT forming a part of the T cell zone, and the coalescent nature of cnB, cnDZ, and cnLZ forming a follicle. These are both visible in Figure 2A (orange/purple/red regions and green/cyan/brown regions). Thus, the different types of local interaction structure indicate that there are differences in the mechanisms by which signals are generated and propagated through the CNs of the T cell zone and follicle.

The barycentric coordinate projections showed that SCs formed by CNs could be described by co-localized combinations of CNs, i.e., the vertices and edges of the triangle which had many cells assigned. We extended this approach to describe the SCs in HLT formed by all the CNs versus only three at a time (as we had in Figures 2B and 2C).

We assigned each cell to the minimal combination of CNs, such that 90% of the cells within its 100 nearest neighbors were assigned to those CNs. This corresponds to explaining the major CNs represented in a window (and thus the major sources of signals to that cell), using as few CNs as possible and the 90% threshold was selected to limit the complexity of the resulting SCs. Thus, this approach corresponds to the generalization of assigning cells to edges and vertices of the triangle in the case of three CNs (i.e., assigning cells to faces of a simplex, the higher-dimensional analog of a triangle, whose vertices each corresponded to a CN). For example, if a cell's window had 80 cells of cnX, 15 of cnY, and 5 of the remaining CNs, it would be assigned to the combination (cnX and cnY). Neither which SCs were identified, or their relative abundances was affected by small variations in the window size or the threshold (Figures S4A and S4B) showing that this approach was a suitable strategy for describing SCs of a tissue despite the parameter choices.

We visualized the follicular and T cell zone regions' abundant SCs (Figures 2B and 2C) as graphs that we termed CNMs (Figure 2D, non-inset). A CNM summarizes the information present in the barycentric coordinate plots and enables visualization and quantification of additional key features. Each node of a CNM corresponds to the SC defined by co-localization of exactly the CNs indicated by the node's label (Figure 2D, inset, single and stacked squares). The size of the node reflects the number of cells assigned to that SC (Figure 2D, inset, black circles). An edge in the CNM indicates that the combination of CNs at the child node is obtained by adding one CN to the combination of CNs of the parent (Figure 2D, inset, vertical arrow). Therefore, the CNMs of the follicular and T cell zone regions explicitly visualize the SCs that are abundant (Figures 2E and 2F, nodes), the CNs participating in each SC (Figures 2E and 2F, colored stacks of squares at each vertex), the number of co-localizing CNs (Figures 2E and 2F, row number and number of squares in stack), and the abundance of each SC (Figures 2E and 2F, size of nodes). The "BT-T-V Intermixing" SC is in the same row as the "B-T Interface" SC rather than on a row below due to limitations on figure space (Figure 2F). The fact that cnDZ-cnB is not included in Figure 2E, and cnBT alone, cnV alone, cnBT-cnV, and cnT-cnV are not included in Figure 2F indicates that there

Cell Systems
Article

CellPress

are not many cells present in those combinations. CNMs therefore provide a description of the SCs present in a tissue.

The CNM for the combined HLT samples and the eleven CNs is presented in Figure 2G with those SCs containing more than 0.001% of total cells shown. The colored groupings of columns and nodes are manually added to illustrate aspects of it that we presently describe. In the following, we denote the SC with, for example, CNs A, B, and C as SC(A, B, and C). Different graphical structures present in the CNM provide different insights into the nature of and interactions between the characteristic local processes of the CNs of HLT.

First, CNs can be found restricted, together, to regions of the CNM, with a small number of edges connecting them with the rest of the CNM. For example, cnLZ and cnDZ (Figure 2G, Column 10) and cnGr and cnMΦ (Figure 2G, Columns 1–2). Such CNs being restricted together suggests that their role in tissue outputs is achieved through signals propagating primarily among themselves. Conversely, other CNs are found in SCs with many other CNs, for example, cnV and cnBT (Figure 2G, purple and orange squares, respectively), and so signal propagation to or from many other CNs could be important for their output. This is consistent with cnV (blood vessels) supplying cells with oxygen or metabolites or enabling cell entry/exit into or from the vascular system.

Second, the tissue was comprised of abundant compartmentalized regions of some CNs (i.e., SCs containing only one CN), and abundant SCs of CNs whose compartmentalized regions were not abundant. For example, cnT and cnB had large, compartmentalized regions (Figure 2G, first row, red and brown squares, respectively), whereas cnPl and cnBT had small, compartmentalized regions but were albeit abundant across the tissue (Figure 2G, gray and orange squares, respectively). A CN having a large, compartmentalized region could be explained by its contribution to tissue output depending on the isolation of its local processes from signals outside of it, either to enrich internal interactions or to avoid interference from external signals. Likewise, a CN with a small (relative to other SCs that CN belongs to), compartmentalized region that is albeit abundant could indicate that its contribution to tissue output does not require isolation of its local processes to the same extent.

Thirdly, the CNM highlights examples where the presence of a CN in SCs with other CNs necessitates the presence of another. For example, CNs T and BT were commonly co-localized together as part of different SCs. SCs were present involving these two CNs co-localized with all other CNs individually, except cnLZ and cnEp (Figure 2G, orange boxes around SCs). The three abundant SCs in which cnT co-localizes without cnBT are SC(V,T) (column 4, row 2), SC(T,MS) (column 4, row 2), and SC(T,Gr, MΦ)) (column 1, row 3) (Figure 2G, red boxes around SCs). However, these three SCs are less abundant than the corresponding combinations in which cnBT is present: SC(BT,T,V) (column 4, row 3,), SC(BT,T,MS) (column 4, row 3), and SC(BT,T,Gr,MΦ) (column 1, row 4) (Figure 2G). This analysis captures the fact that B cells are typically found in SCs where both cells of cnT and other CNs are present. In our biological model, this means both that local interactions between the processes of cnT and other CNs could play a role in recruiting B cells and that signals propagating between cnT and other CNs could be transformed by B cells. Thus, specific relationships between

CNs in how they form SCs are indicative of the underlying mechanisms by which signals are generated and propagated between CNs.

Finally, the CNM highlighted certain CN combinations that were unexpected. For example, we noticed combinations involving cnBT and cnPl with both cnB and cnDZ or both cnB and cnLZ (Figure 2G, black boxes around SCs, column 9, row 4). This was unexpected because we had not initially noticed cnBT or cnPl in the germinal center when visualizing the CNs on the tissues. However, upon closer inspection, these SCs made up a small but definitive region at the interface between cnLZ and cnB (Figure 2H.1, black points). In some follicles, there was an increased frequency of CD38 expression (Figure 2H.2, white staining) and PD-1 expression (Figure 2H.3, green staining) in this region. Upon closer inspection of a representative area (Figures 2H.1–2H.3, white boxes) from this region, the PD-1 expression was restricted to $CD4^+$ cells (Figure 2H.4, box #1, column #1), and the CD38 expression was restricted to $CD19^+$ B Cells (Figure 2H.4, box #2; column #2). In other follicles, these SCs were present (Figure 2I, black points, top row) but were not enriched in $CD38^+$ staining (Figure 2I, bottom row). A thin gap between the germinal center and the mantle zone present in some follicles has been previously defined as the outer zone (Hardie et al., 1993) (Figure 2J, black points); the outer zone contains finer follicular dendritic cell processes than those of the light zone (Steiniger et al., 2011). Thus, the regions where we observed $CD38^+/CD19^+$ B cells co-localized with cnBT and cnPl, and cnB, cnDZ, or cnLZ correspond to outer zones (Figure 2J). The outer zone has also been reported to contain a reduced frequency of $CD57^+/CD4^+$ T cells relative to those in the light zone, but our images of $CD57^+$ staining were not definitive with respect to this observation (Figure S5). Our observations are consistent with the previously reported finding that the outer zone is present only in a minority of follicles (Steiniger et al., 2011) and is also a niche within the follicle for plasma cells (Steiniger et al., 2020). These data are consistent with the possibilities that the SCs identified from the CNM are dynamic during germinal center development or that they are spatially heterogeneous with respect to the three-dimensional orientation of follicles.

In sum, these findings demonstrate that SCs—complex microenvironments defined by co-localization of CNs—are sites where interactions between the local processes of their constituent CNs could occur and that such interactions might be mapped by unique cellular phenotypes. In addition, these data show that simultaneously assessing which SCs as formed as well as the local interactions occurring within each SC in a CNM facilitates understanding of how the tissue achieves complex outputs.

Modular assembly of CN instances differs across HLT types

What role could the interactions between the local processes of CNs occurring in SCs have in facilitating complex tissue behavior? We reasoned that interactions (if they were occurring) in SCs could correspond to the translation of signals so that they could be propagated from one CN to another. For example, a given molecule or cell type produced in one CN might be physically restricted from moving between CNs, or the cells of the
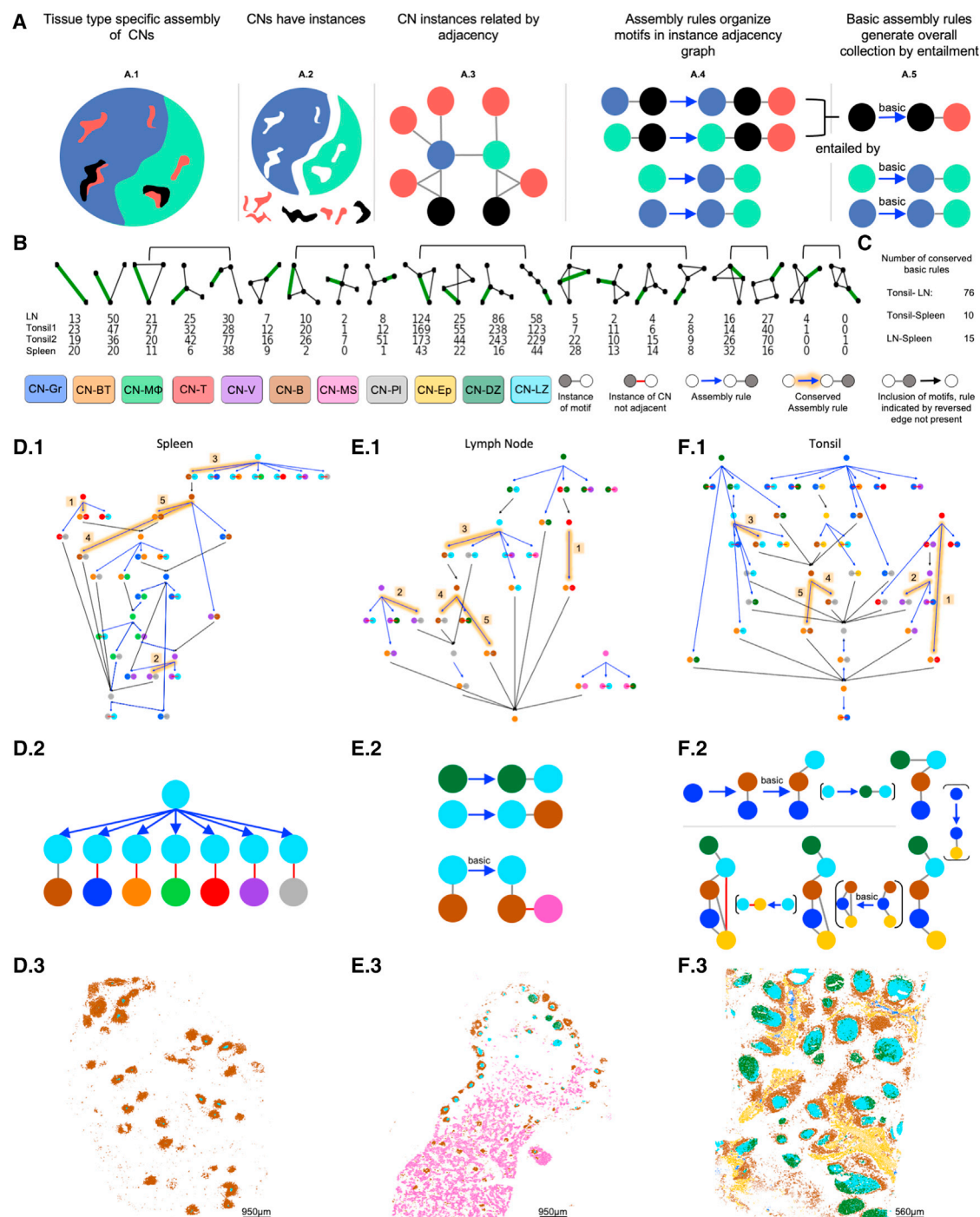
Figure 3. Modular assembly of CN instances differs across HLT

(A) Schematic of analytical approach.

(A.1) Tissues have CNs assembled in different arrangements.

(A.2) CNs have connected components, which are referred to as "instances."

(A.3) Instances in tissue are related by spatial adjacency forming a graph.

(A.4) Motifs in the graph are repeated colored subgraphs and architectural rules correspond to a motif always found as part of a larger one.

(A.5) The overall collection of rules is generated from basic rules by entailment.

(B) Table showing the number of basic rules in each tissue. Basic rules are separated by the topology of the rule. Green edge corresponds to edge added to a motif indicating CNs and adjacencies. Black brackets group rules with same initial motif structure.

(C) The number of basic rules conserved between each pair of tissue types.

(D) Tissue assembly of spleen.

Cell Systems
Article

CellPress

other CN might lack the appropriate receptors to respond to said molecule or cell. However, the signal could trigger the production of a different molecule to which the cells of the receiving CN could respond. Regardless of whether signals were propagated directly between CNs or translated in SCs, we therefore expected that a connected region of one CN could act as a relay, propagating specific signals between other, neighboring CN regions (Figure 1A, right column, cnGreen acting as a relay between cnBlue and cnRed). It would therefore be expected that a collection of connected CN regions, with specific SCs formed between them (which we termed an "assembly"), would be poised to generate unique biological signals by coupling the processes occurring in each individual CN participating in the assembly. If this were the case, there should be an underlying order to how a given tissue type's CNs are assembled, and this order should reflect how relatively simple processes occurring in connected CN regions are chained together to generate complex biological signals.

We developed a systematized approach to explicitly describe the assembly of CNs in tissues such as HLT. We observed that each tissue contained multiple connected components of each CN (and were therefore likely regions through which its characteristic local processes could propagate signals). In an example pseudo-tissue, cnRed is found in five different locations, cnBlack is found in two locations, and cnBlue and cnGreen are each found once (Figure 3A.1). We termed each connected component an "instance." In this model, the CN instances in the tissue are computationally determined (Figure 3A.2). These instances each have their own SCs, and for simplicity, we restricted description of the SCs between instances to pairwise adjacencies (i.e., one instance was either in contact with another, or not). Thus, mapping all the pairwise adjacencies between CN instances results in a graph of the tissue where colored vertices correspond to CN instances and edges correspond to adjacency (Figure 3A.3). We could have incorporated more complex SCs relating instances and that would have resulted in a corresponding tissue hypergraph, but we restricted our attention to the pairwise case. Accordingly, each tissue's graph had instances of "motifs"—repeated subgraphs corresponding to specific combinations of CN instances and spatial relationships. In the example tissue, there are three instances of the cnRed-cnBlue motif and two instances of the cnBlack-cnRed motif (Figure 3A.3). Motifs offered a simple approximation for CN assemblies that required minimal assumptions on signal propagation within a CN and on how CNs were connected in three-dimensional space.

If CN assemblies had been evolutionarily selected, we would expect there to be a collection of "ARs" governing how CN in-

stances are assembled into motifs, that is characteristic of a given tissue type, but also would hold in tissues that carried out similar functions. Moreover, we would expect this evolutionary process to have given rise to modularity within a given tissue type (i.e., ARs dictating instances of a given CN being found in multiple distinct motifs) and across distinct tissue types (i.e., conservation and repurposing of ARs across distinct HLTs). An AR present in the example tissue is that in every instance of cnBlue-cnBlack, cnBlack is adjacent to cnRed (Figure 3A.4, top row). This could indicate that that an instance of cnBlue-cnBlack-cnRed is required for an instance of the cnBlue-cnBlack to develop and would also suggest that adjacency to cnRed is important for the functional role of the cnBlue-cnBlack in the tissue. Another example that is not present in the example tissue is, only when an instance of a CN is adjacent to a second CN, is it adjacent to an instance of a third CN. This rule could indicate that instances of the first and second CNs provide complementary factors that interact to generate an instance of the third. The overall collection of ARs for a tissue type was defined as the collection of rules that specify that an instance of a motif is always part of an instance of a larger motif (Figure 3A.4, four ARs from the overall collection are indicated with blue arrows).

Given the overall collection of ARs, we reasoned that there would be a subset of rules that generated the overall collection of architectural rules and were 'basic,' in the sense that they could not be inferred from simpler rules (Figure 3A.5). For example, the rule introduced above (i.e., in every instance of cnBlue-cnBlack, cnBlack is adjacent to cnRed), is entailed by a simpler rule that cnBlack is always adjacent to cnRed (Figure 3A.5, rule indicated by blue arrow labeled "basic"). By capturing redundancy in the overall collection of ARs, the basic ones provide a parsimonious approach to represent and interrogate the architecture. Moreover, a given basic rule indicates an emergent architectural constraint, i.e., when a motif is always extended in a certain way, but its submotifs are not constrained in that same manner. The basic ARs thus highlight likely examples of motifs that can produce signals in coordinate which its submotifs cannot.

We approximately identified the basic ARs for each type of lymphoid tissue (see STAR Methods) and quantified the number of such rules in each tissue, partitioning them by the type of motif featuring in the rule (Figure 3B, green edge indicates added adjacency for that type of rule). For example, three types of extension are possible from a motif consisting of three CNs (X-Y-Z) in a chain: adjacency between cnX and cnZ, addition of a cnW on to cnY, and lengthening to a chain of length four by adding cnW on to cnZ or cnX (Figure 3B, left most bracket). The spleen had the fewest basic rules involving more than two CNs, whereas

(D.1) Graph of basic rules in spleen. Orange highlighted arrows indicate conserved AR.
(D.2) Basic rules describing topology of CNs LZ and B in spleen.
(D.3) Realization of rules of (D.2) with tissue CN plot for spleen.
(E) Tissue assembly of lymph node.
(E.1) Graph of basic rules in lymph node. Orange highlighted arrows indicate conserved assembly rules.
(E.2) Basic rules describing organization of CNs LZ, DZ, and B in lymph node.
(E.3) Realization of rules of (E.2) in tissue CN plot for lymph node.
(F) Tissue assembly of tonsil.
(F.1) Graph of basic rules in tonsil. Orange highlighted arrows indicate conserved assembly rules.
(F.2) Basic and inferred rules describing context of cnGr in tonsil.
(F.3) Realization of rules of (F.2) in tissue CN plot for tonsil 1.

the two tonsils each had a larger number of such rules (Figure 3B). Under this regime, the spleen therefore has a simpler architecture than tonsil (and potentially participates in fewer immune outputs diversified through repurposing architectural programs). The type of motif with the largest number of basic rules across all tissues were those involving four CNs in a chain (Figure 3B, third bracket from left). Chains minimally constrain the instances within them so there are many rules involving four CNs. We found that 76 basic rules of this nature were shared by tonsil and lymph node, 10 by spleen and tonsil, and 15 by spleen and lymph node. Thus, the basic rules provide automated quantification of how, as has been long described, the tonsil and lymph node are architecturally more similar to each other than either is to the spleen (Figure 3C).

We visualized the rules for each tissue that involved a single CN instance being adjacent to a second CN instance graphically (Figures 3D.1, 3E.1, and 3F.1). We validated that these rules were not sensitive to the exact threshold on the frequencies that each instance of a CN extended to the larger motif used to define them (Figure S6). Such rules are necessarily basic because they relate motifs containing only one and two CNs (and thus cannot be any simpler). The vertices in each graph indicate motifs. Gray and red edges describe properties of the motif at each vertex. Gray edges between CNs indicate an instance of a CN adjacent to an instance of another CN. Red edges between CNs indicate an instance of a CN not adjacent to an instance of another CN and are abbreviated in the text as "cnXnotY." Blue and black edges indicate ARs. Blue edges indicate ARs of the form: "given an instance of the motif at the source of the arrow, there exists an extension of that instance to the motif at the target." We refer to such a rule in text as cnX→(cnX-cnY) (i.e., given an instance of cnX, it is found as part of an instance of the motif with cnX adjacent to cnY). Black edges indicate when the target is a submotif of the source (i.e., (cnX-cnY)→cnX). Therefore, traversing the graph corresponds to modeling more complicated ARs by combining basic ones.

Five basic rules involving two CNs were conserved in tonsil, lymph node, and spleen: (1) cnT→(cnBT-cnT), (2) cnV→(cnV-cnPl), (3) cnLZ→(cnB-cnLZ), (4) cnB→(cnB-cnPl), and (5) cnB→(cnB-cnBT) (Figures 3D.1, 3E.1, and 3F.1, orange highlighted arrows numbered 1–5). These rules are therefore the core lymphoid ARs. The latter three rules entail that cnLZ is always assembled into the larger motif (cnLZ-cnB-(cnPl,cnBT)), where the ","indicates that cnB is adjacent to both cnPl and cnBT in the rule. Moreover, cnB→(cnLZ-cnB) is not a rule in any tissue. This suggests that interactions between signals propagating in cnLZ and those processed by the cnB-(cnBT,cnPl) motif could be essential for the role of cnLZ in lymphoid tissue.

There were also architectural rules unique to each tissue. In the spleen, cnLZ→(cnLZnotX) (wherein "not" is a designator indicating that an instance of LZ is not adjacent to any instance of cnX) is a basic rule for all CNs X present in the spleen with the exception of cnB (Figure 3D.2), meaning that cnLZ instances are all surrounded by cnB (Figure 3D.3). As such, all signals propagating between cnLZ and other CNs pass through and are potentially transformed by the local processes of cnB. This is in contrast with the tonsil and lymph node wherein cnLZ→cnLZ-cnPl is a basic rule (Figures 3E.1 and 3F.1, edges from cnLZ to cnLZ-cnPl in each tissue), indicating that cnLZ instances are not surrounded only by cnB in these tissues.

In both the tonsil and lymph node cnDZ→cnLZ-cnDZ is an architectural rule. However, cnLZ→cnDZ-cnLZ is not a rule in the lymph node. In other words, every germinal center in the tonsil has cnDZ-cnLZ, whereas in the lymph node only a subset of them have cnDZ (Figures S3E and 3F). These differences in the rules could be attributed to differences in tissue type between lymph node and tonsil, or to the individual, activation state, or 3D orientation of the individual specimens in our dataset. In addition, in the lymph node, we found a basic rule cnB-cnLZ→cnLZ-cnBnotMS (Figure 3E.2), meaning that while there were instances of cnB adjacent to cnMS, there were no instances of cnB adjacent to both cnMS and cnLZ. These basic rules correspond to two kinds of cnB instances in the lymph node: those in the paracortex adjacent to cnLZ (i.e., mantle zone instances), and those in the medullary sinuses that are not adjacent to cnLZ (Figure 3E.3). However, in both cases, these instances of cnB were adjacent to instances of cnBT and cnPl since this was ensured by the basic rules across lymphoid tissues. Thus, these basic rules suggest that cnB-(cnBT,cnPl) is a modular component playing a similar role in both the paracortex and medullary sinus, with the signals that facilitate the development of the larger assembly with cnLZ present only in the paracortex.

In the tonsil, we inferred an architectural rule that demonstrates the assembly of a follicular structure that involves cnGr, wherein cnLZ and cnDZ are separated from cnEp by cnB. (Figure 3F.2, begin at cnGr in top row, farthest left, and combine rules to the right, down to bottom row, and back to left). This could indicate that either the complex motif indicated is necessary for the development of cnGr to occur, or that the signals generated and processed in cnGr must be processed by the other CNs of the motif to participate in tonsillar function. These data also indicate that cnB provides a physical bridge from within the mucosa of the epithelium to the mantle zone of the follicle (Figure S7). The B cells in the subregion of cnB within the mucosa of the epithelium are likely enriched in memory B cells as previously reported (Liu et al., 1995). This suggests that CNs could be differentiated, with respect to their functional states (as defined in Schürch et al., 2020) in specific locations to enhance signal propagation between other components of complex assemblies.

Mapping the wiring and interactions between the local processes of the CNs of the colorectal cancer iTME

We turned to a distinct CODEX dataset of the tumor microenvironment (iTME) in human colorectal cancer (CRC) (Schürch et al., 2020) to determine if the approach used for healthy lymphoid tissue could be applied to diseased tissue that also contained lymphoid cells and incipient lymphoid structures. The data came from a study that had compared the iTMEs of two previously defined groups in the colorectal cancer spectrum: the "CLR" group (17 patients) (Graham and Appelman, 1990) and "diffuse inflammatory infiltration" (DII) group. The CLR group is classically defined by the presence of TLSs (Figure 4A.1, top left illustration, containing blue dots representing TLS), and patients in this group characteristically have good survival outcomes as compared with those in the DII group, defined by their lack of TLS (Figure 4A.1, left illustration, no blue dots) that have poor survival outcomes. This study had compiled four 0.6-mm diameter tissue cores from each patient onto two 70-core next
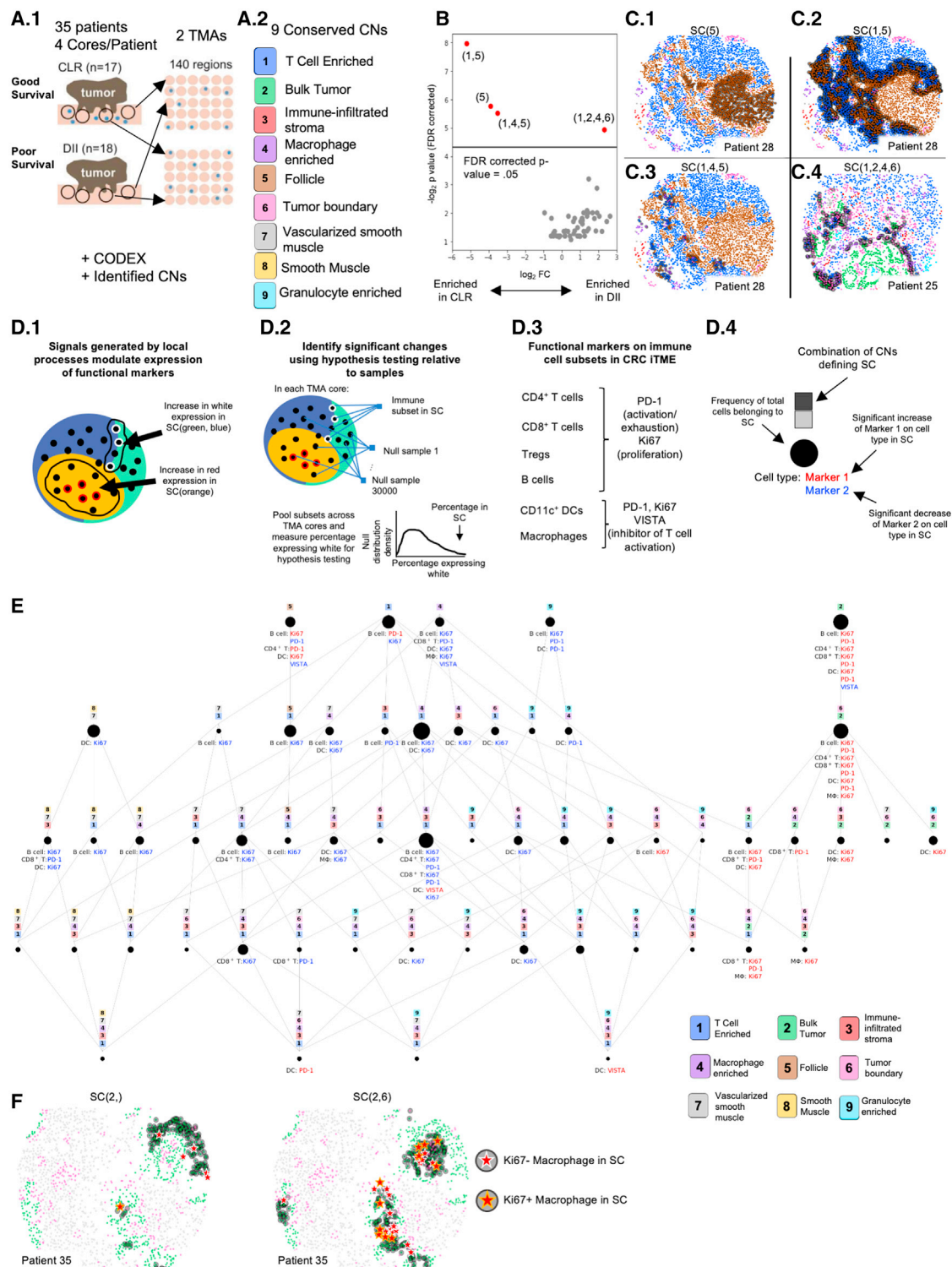
Figure 4. Mapping the interactions between the local processes of the CRC iTME

(A) (A.1) Schematic of CRC dataset, consisting of four TMA cores imaged from each of 17 CLR patients who had TLS (blue dots) and good survival outcomes and 18 DII patients, who had no TLS and poor survival outcomes. Reproduced from (Schürch et al., 2020) under a CC-BY 4.0 license. (A.2) CNs that had been previously identified and validated.

(B) Volcano plot showing CN combination SCs that were present in at least 10 patients and were differentially abundant in one patient group as determined via Mann Whitney U tests for differences in the frequency of each SC. y axis indicates $-\log_2$ Benjamini-Hochberg FDR. Line indicates FDR = 0.05.

(C) Visualizations of differentially abundant SCs in individual TMA cores. (C.1): SC(5), (C.2): SC(1,5), (C.3): SC(1,4,5), (C.4): SC(1,2,4,6).

(legend continued on next page)

generation tissue microarrays (TMAs), enabling the analysis of this large number of patients (Figure 4A.1, right illustration [which was obtained from Figure 3A in the original publication; Schürch et al., 2020]). The clear connection between the presence of the follicular structure and patient survival alongside the large sample size in this dataset also afforded the opportunity not just to construct a TS of the CRC iTME but also to dissect how lymphoid architecture is deployed in newly arisen peripheral sites to combat tumors.

The first ingredient of a TS—the identification of CNs and assessment of their possible local processes—had been previously performed. The previous study had identified 9 CNs in the CRC iTME, all of which were conserved between CLR and DII patients (Figure 4A.2), and all of these were present at similar frequencies in both CLR and DII patient groups except for cn5, corresponding to the TLS. Moreover, the functional states of the iTME's CNs were defined and shown to be correlated, and these correlations differed between high and lower-risk patient groups, indicating that the local processes of each CN interacted and were implicated in antitumoral immunity (Schürch et al., 2020). However, the role that these local processes and their interactions in SCs had in antitumoral immunity was not previously mapped at the granularity proposed here.

The SCs formed by the CNs of the CRC iTME could be differentially abundant between patient groups. If so, this could suggest the importance of such an SC in the more effective immune response observed in CLR patients. We identified SCs using a window size of 50 nearest neighbors and a threshold of 90% when allocating CNs (see STAR Methods). We only assessed SCs that were conserved across multiple patients, selecting those SCs with at least 25 cells in at least 10 patients, resulting in 55 SCs. Next, we identified SCs that were significantly enriched in one patient group by computing the overall frequency of each SC in each patient (pooling together cells from each of their 4 TMA cores) and conducting a Mann-Whitney U test for differences. Four SCs were significantly different in their abundance between the two groups after adjusting for multiple hypothesis testing (FDR < 0.05) (Figure 4B, labeled red points). SC(5), corresponding to the compartmentalized region of the follicle (Figure 4C.1), SC(1,5) corresponding to an interface between cn1 (T cell enriched) and cn5 (follicle) (Figure 4C.2) as well as SC(1,4,5) corresponding to three-way combination of cn1 (T cell enriched), cn4 (macrophage enriched), and cn5 (follicle) (Figure 4C.3) were enriched in CLR patients. SC(1,2,4,6) corresponding to the four-way combination of cn1(T cell enriched), cn2(bulk tumor), cn4 (macrophage enriched), and cn6(tumor

boundary) (Figure 4C.4) was enriched in DII patients. The previous study had reported the possibility of communication and other interactions between the cellular compositions of CNs 1, 2, 4, and 6, and thus, this communication could be associated with direct spatial contact.

What signals are generated as a result of interactions between the local processes of CNs in the SCs that are formed? We reasoned that signals generated by the interactions between the local processes of each CN in any given SC would be manifest as a shift in the states of the immune cells found therein, relative to those found in iTME as a whole (Figure 4D.1, in the diagram, a large proportion of black cells in SC(blue,green) express the white marker and a large proportion of black cells in SC(orange) express the red marker). The major immune subsets (CD4$^+$ T cells, CD8$^+$ T cells, B cells, and macrophages) in the CRC dataset had been previously gated for expression of functional immune markers: PD-1 (activation/exhaustion), Ki67 (proliferation), VISTA (immune checkpoint expressed on myeloid cells suppressing T cell activation), and these markers were not used in the definitions of cell types, or identification of CNs (Figure 4D.3). Given this, we identified examples of immune cells whose expression of a marker was significantly higher (or lower) in each SC relative to the rest of the tissue (Figure 4D.2). Specifically, for each SC, immune cell subset and functional marker, we constructed a null distribution by sampling 30,000 random subsets of immune cells of the same size as the SC, fixing the number from each TMA core and computing the frequency of cells positive for each functional marker (Figure 4D.2, selected null samples as well as SC sample illustrated by groupings of black cells). Next, we compared the frequency of cells in the SC that were positive for each marker with the corresponding null distribution to obtain a p value (Figure 4D.2, distribution plot in lower right).

This approach had the advantage that no information about the expression of markers leaked into the CN or cell-type allocation, so significant results indicate a genuine biological association between a given SC and the state of an immune subset. However, the approach cannot distinguish whether migrating cells modulate their expression upon entering these sites or new subsets preferentially localize there (potentially through trafficking biases or increased proliferation/death upon reaching the SC). The CNM is presented in Figure 4E. The red color of the font in the cell subset and functional markers below each node (identified at an FDR of 0.05) indicate that the proportion of that subset expressing the marker is higher in the SC relative to the null distribution, and blue correspondingly lower.

(D) Schematic to illustrate approach for identifying mapping effects of and interactions between local processes of CNs.

(D.1) A schematic tissue with three CNs, their SCs, and a cell type (black) illustrated. The cell type can express the white and red functional markers. Circled regions indicate SC(orange) and SC(green, blue).

(D.2) A sample from the null distribution is obtained by in each TMA core selecting the same number of cells as are found in the given SC, pooling the cells across cores and computing the percentage of those cells that are positive for a given functional marker. These samples are compared with the percentage positive in the SC to compute a p value.

(D.3) The immune subsets and functional markers available in the CRC dataset.

(D.4) Legend for CNM, visualizing SCs where a cell subset has an increased percentage expressing a functional marker are listed in red, and a decreased percentage in blue.

(E) CNM for CRC iTME annotated with changes in expression of functional markers on key immune subsets. The p values for tests for combinations of cell types and markers were combined and Benjamini-Hochberg FDRs were computed. Displayed changes are FDR < 0.05.

(F) Plots of SCs and macrophage phenotypes on selected TMA cores. For clarity only CNs 2 and 6 are displayed. SCs in their respective plots are backlit in dark gray. Stars with white outlines indicate Ki67− macrophages and stars with orange outlines indicate Ki67+ macrophages.

Cell Systems
Article

CellPress

We first observed that the local processes of the follicle recapitulated those expected to be in a healthy lymphoid organ. In the compartmentalized follicle SC(5), we observed a significant increase of proliferation on B cells and DCs, and a reduction of VISTA on DCs and a decrease of PD-1 on B cells (Figure 4E, annotations below leftmost node in top row). In addition, SC(5) was the only CN where there was an increase of PD-1 expression on CD4$^+$ T cells, consistent with the presence of PD-1+ T follicular helper cells in the TLS. From this, it can be inferred that within the follicle of the CRC iTME, the local processes involve antigen driven activation driven interactions between these subsets. Since the only descendant of SC(5) was SC(1,5) (Figure 4E, edge from SC(5)), the follicle could be deduced to be generally surrounded by cn1. In SC(1), there was an increase of PD-1 expression on B cells, along with a reduction of proliferation. These results indicate an analogous wiring of the iTME TLS with lymphoid organs, where B cells are activated in the T cell zone prior to forming follicles (Nutt and Tarlinton, 2011).

As with the follicle there was an increase in the compartmentalized part of the tumor, SC(2), of DC and B cell proliferation and a reduction of DC VISTA expression (Figure 4E, rightmost node). However, there was also an increase in proliferation of CD8$^+$ T cells, and PD-1 expression on both CD8$^+$ T cells and B cells. This indicates that the local processes of the tumor in addition to the follicle give rise to activation of immune cells that could be antigen dependent, but with a different milieu (not giving rise to the activated CD4$^+$ T cells or reduction of PD-1 on B cells). As expected from the previous study, we noticed that the tumor was also generally surrounded by cn6 (tumor boundary, Figure 4E, only descendant of SC(2)). We noticed that proliferation of macrophages was increased in SC(2,6) as well as SC(2,3,6), SC(1,2,4,6), and SC(2,3,4,6), but not in those SCs containing cn6 that do not also contain cn2 (Figure 4E, compare nodes annotated with macrophage Ki67 with nodes with green and pink) or in SC(2). Thus, the local processes of cn2 (bulk tumor) and cn6 (tumor boundary) could interact to generate signals that increase macrophage proliferation (Figure 4E, red Ki67 for macrophages). While this result was obtained pooling the data from across patient cores, we depicted the macrophages expressing and not expressing Ki67, in SC(2) and SC(2,6) in one TMA core to illustrate the finding (Figure 4F).

We next examined the local processes of cn4 (macrophage enriched). Its compartmentalized region SC(4) (Figure 4E, top row, third from left) was the only SC which had a decrease in VISTA expression on macrophages, indicating that the local processes therein might generate unique signals for regulation of macrophage phenotypes. SC(1,3,4) had a reduction in proliferation of DCs, B cells, CD4$^+$ T cells, and CD8$^+$ T cells, a reduction in expression of PD-1 on CD4$^+$ and CD8$^+$ T cells (Figure 4E, node in third row with many blue annotations). Moreover, SC(1,3,4) and SC(1,3,4,6,9) were the only SCs in which an increase in the expression of the inhibitory marker VISTA on DCs was observed. This indicates that the local processes of cn1 (T cell enriched), cn3 (immune-infiltrated stroma), and cn4 (macrophage enriched) interact in SC(1,3,4) to generate signals leading to the inhibition of T cells that could be mediated by DCs. In SC(1,2,4,6), there was increased PD-1 expression on CD8 T cells and proliferation of CD8 T cells and macrophages (Figure 4E, bottom row, second

from right), and so, the local processes of cn4 alone are not sufficient for such inhibitory signals.

It was previously shown (Schürch et al., 2020) that an increased frequency of CD4$^+$ PD-1+ T cells in cn9 (granulocyte enriched) was associated with improved survival outcomes in DII patients. Whereas there was an increase in PD-1 expression on DCs in three SCs involving the tumor: SC(2) and SC(2,6), as well as the combination of the inhibitory combination of CNs along with the tumor SC(1,3,4,6,7), there was a reduction in PD-1 expression on DCs in SC(9) and SC(4,9). Therefore, the local processes of cn9 (granulocyte enriched) and cn2 (tumor) have opposing associations with DC phenotypes that could play a role in the antitumoral immune response.

Thus, identifying the SCs formed by the CNs of the CRC iTME, as well as the changes in expression of functional markers on key immune subsets is informative of the wiring of the iTME.

Higher-order tissue assembly orchestrates antitumoral immunity in CRC

The CNM for the CRC iTME mapped the local processes of each CN of the iTME and how they interacted as pertained to antitumoral immunity. Our results suggested that, as was the case in HLT, their CNs would form instances that could assemble into motifs with emergent biological functions. Are these motifs mere statistical consequences of interactions that are essentially simpler to describe? Or are these motifs actively assembled by the iTME in a way that drives patient outcomes?

As we had done for HLT, we segmented instances of each CN and constructed a tissue graph for each TMA core (Figure 5A), considering only instances of any CN with more than 5 cells (see STAR Methods). We observed that the tissues of DII patients were significantly more fragmented than those of CLR patients, as measured by the average number of CN instances (vertices) in the tissue graphs for each patient (p = 0.006) (Figure 5B, left), as well as the average degree (number of edges emanating from a vertex) (p = 0.02) (Figure 5B, right). This raised the question: did the fragmented tissue graphs of the DII patients mean that the assembly of CNs into motifs was simply in a state of disorder with individual CNs unable to assemble into the motifs required for emergent tissue behaviors, or did the assembly correspond to an alternative state of tumor-favoring order?

We began by computing the frequency that each instance of the tissue graph was found as part of a chain of 2 CNs (hereafter referred to as a two-chain) and identified candidate ARs as we had done with the TS for HLT in each patient group, visualizing combinations where this frequency was above 0.65 in either patient group (Figure 5C, color indicating frequency of extension). We found that most of the candidate rules were common to both patient groups, for example bulk tumor (cn2) instances were always extended to the cn2–cn6 (bulk tumor/tumor boundary) motif (Figure 5C, top row), smooth muscle instances were always extended to the cn7-cn8 (smooth muscle/vascularized smooth muscle) motif (Figure 5C, second row from top). The identified cell types were qualitatively comparable between the HLT and CRC datasets, and so we could approximately compare certain CNs (cn1 could be mapped to either cnBT or cnT, cn5 mapped to cnB, cn4 mapped to cnMΦ, and cn7 mapped to cnV), (Figure S8). Thus, we could map the candidate rules
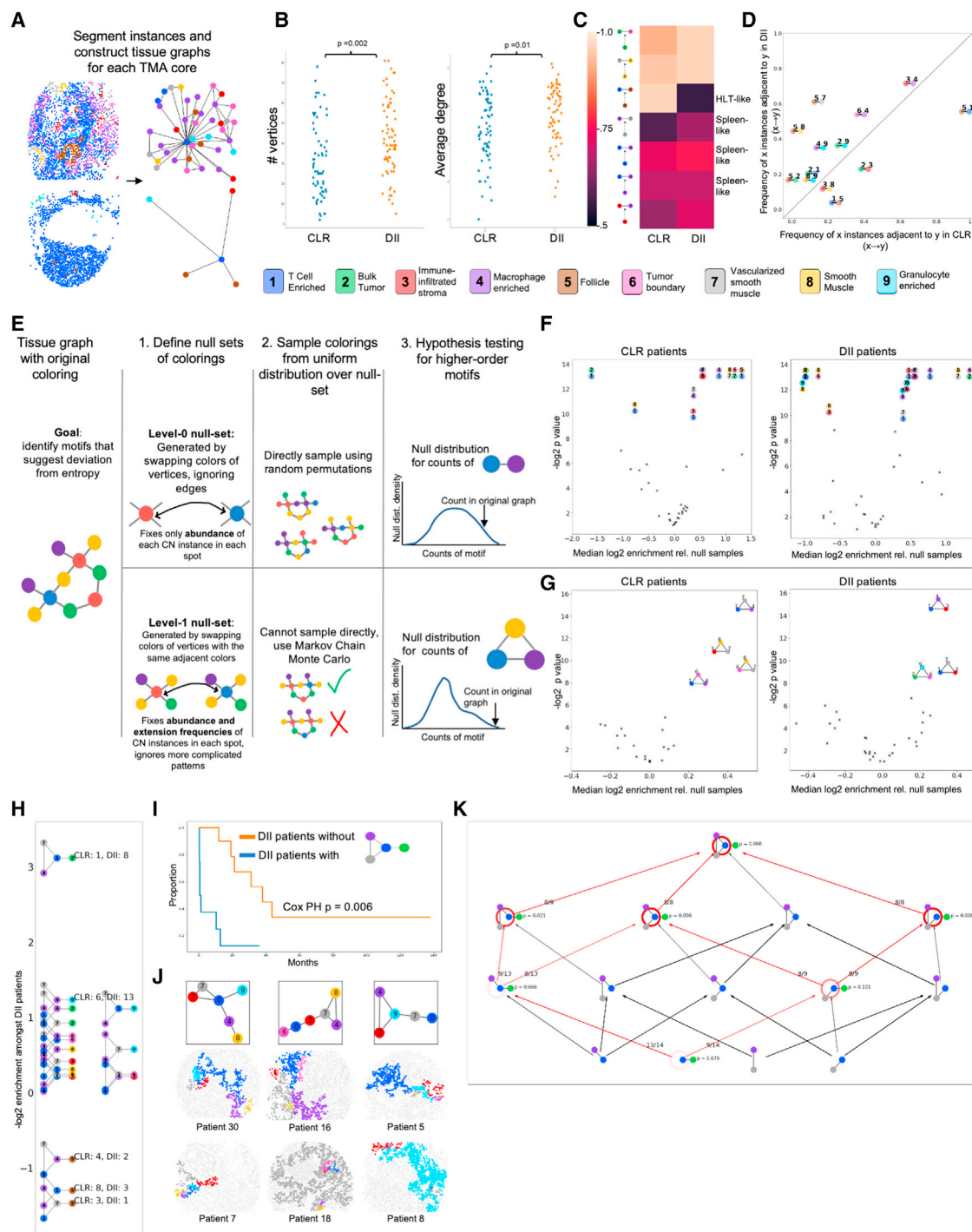
Figure 5. Higher-order tissue assembly orchestrates antitumoral immunity in the CRC iTME

(A) Example tissue graphs constructed from TMA cores by segmenting instances and connecting those that are adjacent with edges.

(B) Quantification of fragmentation of tissue in CLR and DII patients by left: number of vertices and right: average edge degree in the tissue graph of each TMA core, separated by patient group. Mann-Whitney U test for differences in p value applied after the counts across the four cores for each patient had been averaged.

(C) Heatmap of extension-frequencies for instances of lower CN at source of arrow on each row being extended to two-chain at target. Color indicates frequency of instances that do extend, according to scale. Labeled rows indicate approximate correspondences with HLT basic rules.

(D) Motif extensions whose frequencies were significantly different (p < 0.05 permutation test, Bonferroni corrected) between patient groups. The annotated cnA→cnB is an abbreviation that denotes extension frequency of instances of cnA being extended to a two-chain cnA-cnB.

(E) Schematic of identifying higher-order structure relative to a maximum-entropy null distribution over colorings of tissue graph. Top row, from left: A null set is constructed corresponding to set of all colorings with the same number of vertices of each given color, which can be viewed as generated by applying arbitrary

(legend continued on next page)

Cell Systems
Article

CellPress

observed in CRC to those present in the various HLTs (Figure 5C, annotations on right).

We found two rules that were present in only the spleen that were also rules in CRC, corresponding to cn1 (T cell enriched) and cn4 (macrophage enriched) each being extended to the cn1-cn4 motif (Figure 5C, 5th and 6th row from top). In CLR patients, there was a rule: cn5→cn1-cn5. The exact analog of this under the mapping would be cnB→cnB-cnT, which was not a core lymphoid rule. Instead, the core lymphoid rule was cnB→cnB-cnBT (Figure 5C, 3rd row from top). These data indicate that in the CRC iTME, unlike in lymphoid tissue, the TLS assembles to directly contact the T cell-enriched CN, without assembling next to an intermediate CN in which B cells and T cells mix. At the same time, one of the splenic rules, cnV→ cnMΦ-cnV was a rule only in DII patients. Thus, the similarities between the architecture of HLT and the CRC iTME extended beyond just the existence of the follicle, but also included ARs involving the T cell-enriched CNs and the macrophage-enriched CNs.

We assessed whether the differences in rules that were observed in Figure 5C were statistically significant, by permuting patient group assignments prior to computing frequencies of each extension and thus obtaining a null distribution which we could identify significant differences between patient groups (see STAR Methods). This confirmed that cn5 (follicle)→cn1 (T cell enriched)-cn5 (follicle) (analog of the core lymphoid rule of cnB→cnB-cnBT) was only a rule in CLR patients (Figure 5D, indicated by the cn5→cn1 marker on the plot). In DII patients, cn5 (follicle) instances extended to a motif with cn8 (smooth muscle), or cn7 (vasculature), or cn2 (bulk tumor) significantly more frequently in DII patients (Figure 5D, motifs above the diagonal with cn5 on left). We observed that in cn2 (bulk tumor) instances were extended to the cn2 (bulk tumor)-cn9 (granulocyte enriched) and cn2 (bulk tumor)-cn1 (T cell enriched) motifs in DII patients, but to the cn2 (bulk tumor)-cn3 (immune infiltrated stroma) in DII patients (Figure 5D, motifs with cn2 on left).

The fact that both patient groups exhibited ARs that were analogs of those found in the spleen, but only the CLR patients deployed the core lymphoid AR (cnB→cnB-cnBT) suggests that there were common programs of tissue assembly at play in

both patient groups but that were altered between CLR and DII patients. Could these alterations play a role in patient survival outcomes? Given the changes in the extension frequencies involving tumor instances between patient groups, could the tumor play a role in altering the assembly of the CLR patients' iTME?

We quantitatively described the order present in the assemblies of each patient groups' iTMEs. We identified motifs in each patient group that could be considered "higher-order" in the following sense: those, for which there was statistical evidence that the iTME was actively expending energy to assemble. We reasoned that those motifs would be the ones for which there was evidence of deviation from what would be entropically expected, i.e., those whose observed count deviated from a null distribution with high entropy, given some constraints. For example, tissues might actively constrain just the number of each CN instance, but the two-chains formed by them might be expected from a high-entropy distribution given those constraints (and so, in that case, the two-chains would not be higher-order motifs). Alternatively, tissues might actively constrain the numbers of instances of CNs and two-chains, but more complex motifs might be expected by chance. We would expect that the motifs that were higher order, since they are likely the ones that are actively assembled (requiring energy to counteract entropy), would be the ones that could have important roles in tissue outputs that are not accomplished by their subcomponents. Therefore, if DII patients were simply more disordered, we would expect there to be fewer higher-order motifs in their tissue graphs.

We identified higher-order motifs by constructing a null distribution of tissue graphs that was maximally entropic, once some level of "lower-order" structure among these graphs had been fixed (Figure 5E, overall panel). Motifs whose count in the original tissue graph significantly deviated from such a maximum entropy null distribution would indicate higher-order motifs being actively assembled (Figure 5E, column headings). We tested our approach to find two-chains that were higher-order relative to the maximum entropy null distribution obtained when the underlying graph structure and total number of each CN instance in each TMA core had been fixed. This maximum entropy null distribution corresponds to the uniform distribution on the set of

transpositions to the original coloring. The uniform distribution over this set is directly sampled using random permutations. The counts of two-chains with respect to the original coloring are compared with the counts of two-chains in the original coloring to identify significantly deviant ones. Bottom row, from left: instead of a maximum-entropy null distribution that fixes only the number of vertices assigned to each color, the level-1 null-set additionally constrains extension frequencies and is generated applying transpositions only between colors of vertices whose neighboring vertices have the same set of colors. The uniform distribution over this set cannot be directly sampled, so a Monte-Carlo Markov-Chain algorithm is used to sample colorings. The counts of triangles are compared with the counts of triangles in this null set, using hypothesis testing to identify higher-order motifs.

(F) Two-chains whose count significantly deviates from the level-0 maximum-entropy null distribution in CLR patients (left) or DII patients (right) indicated by colored motifs on volcano plots.

(G) Triangle motifs (colored) whose count significantly deviates from the level-1 maximum-entropy null distribution in CLR (left) and DII (right) patients. In both plots, x marker indicates not-significant motifs and colored motifs indicate those significant at FDR = 0.05.

(H) Log$_2$ enrichment of the extensions of the (T cell enriched, macrophage enriched, and vasculature) triangle in DII patients relative to CLR patients. Numerical annotations for enrichment whose absolute value is greater than 1 indicate number of patients with that extension in each patient group.

(I) Kaplan-Meier curves for DII patients with and without the tumor (T-cell-enriched, macrophage-enriched, and vasculature) motif. Cox proportional-hazards regression p value = 0.006.

(J) Examples of motifs present only in patients that do not have the tumor (T-cell-enriched, macrophage-enriched, and vasculature) motif (top row) visualized in the corresponding tissue graphs (bottom row).

(K) Submotifs of the tumor (T-cell-enriched, macrophage-enriched, and vasculature) motif arranged by extension of motifs. Color intensity of circle indicates Cox proportional hazards p value for association with survival in DII patients. Edge weights and labels indicate number of DII patients with motif at source that also have motif at target. Red edge color indicates tumor incorporation into the assembly of the (T-cell-enriched, macrophage-enriched, and vasculature) triangle.

colorings of the original tissue graph with a fixed number of each color, which we refer to as the Level-0 null set and can be viewed as being generated by repeated swaps of the assignment of colors (Figure 5E.1, top row) in the tissue graph within a given TMA core. Samples from the maximum entropy null distribution can be obtained by random sampling of permutations of CN assignments of the vertices for each core (Figure 5E, column 2), and the count of a given two-chain in the original graph can be compared with the count of that two-chain in the colored graphs sampled from the null distribution (Figure 5E, column 3). The two-chains that significantly deviated from the maximum entropy null distribution after Bonferroni correction are visualized in each patient group (Figure 5F, top row, colored motifs). These show that many of the higher-order two-chains are common to both CLR and DII patients, including cn2-cn6 (tumor/tumor boundary), cn7-cn8 (smooth muscle/vascularized smooth muscle and cn1-cn4 (T cell enriched/macrophage enriched).

Instances of cn1–cn2 (T cell enriched/bulk tumor) are less frequent than expected under the maximum entropy null distribution, which can be interpreted as evidence that active tissue processes prevent contact between these two CNs. In addition, while the cn9 (granulocyte enriched) is present in both patient groups at similar frequencies (Schürch et al., 2020), higher-order two-chains involving it are found only in DII patients. In DII patients, two-chains involving cn8 (smooth muscle) with CNs 1, 3, 6, and 9 (T cell enriched, immune-infiltrated stroma, tumor boundary, and granulocyte enriched, respectively) are depleted, further adding to the evidence that some active processes could be constraining the assembly of smooth muscle CNs adjacent to the tumor boundary. It was previously reported that the functional state of cn9 (as measured by the frequency of PD-1+CD-4+ T cells) was associated with survival in DII patients (Schürch et al., 2020), while the analysis of the SCs showed that the local processes of cn9 could be playing a role in counteracting effects the tumor had on DC phenotypes (Figure 4E). The fact that there are significant two-chains involving cn9 in DII patients indicate that its assembly into the iTME is actively specified, and thus adjacencies between its instances could play a role in its antitumoral effect. Taken together, these data did not suggest that the DII patients were simply more disordered, but instead that there was a different order being imposed.

Although both patient groups had similar numbers of higher-order two-chains, we reasoned that perhaps CLR patients would have more triangle motifs (three instances all connected) that were higher-order relative to a maximum entropy distribution obtained by constraining the number of 2-chains. Whereas the level-0 null set, fixing just the number of each CN instances, was generated by arbitrary swaps of vertex CN assignments from the original graph of each spot (Figure 5E1, upper column), we defined the level-1 null set, generated by allowing swaps of the CN assignments of two instances in the tissue graph of a TMA core only if the set of CNs of their neighboring instances is the same (Figure 5E.1, lower column, see STAR Methods). The level-1 null set consists of colorings, which have the same counts of two-chains as the original graph, but the counts of more complex motifs vary. Triangles enriched relative to the uniform distribution on the level-1 null set are likely actively assembled by the iTME and thus could play an active role in antitumoral immunity. While this maximum entropy null distribution cannot

be directly sampled because of the complicated constraints on swaps, we developed a Markov Chain Monte Carlo algorithm to approximately sample from it (see STAR Methods), and we used these samples to identify higher-order triangles. Since these null distributions and the level-0 ones were generated relative to the starting graph, they made minimal prior assumptions on its structure such as CN abundances, number of instances for each CN, or number of outgoing edges.

Triangle (1,4,6) (T cell enriched, macrophage enriched, and tumor boundary) was present significantly more often than in the maximum-entropy null distribution only in CLR patients (Figure 5G, left, colored triangle). Given our results from the CNM (Figure 4E) that the local processes of cn6 (tumor boundary) in combination with those of cn2 (bulk tumor) generates signals for macrophage proliferation, and SC(1,2,4,6) is linked to the activation of CD8+ T cells, the fact that this triangle is actively assembled only in CLR patients indicates that formation of instances of these CNs in direct contact with each other and the tumor could be an active component of the antitumoral response. In contrast, triangle (1,3,4) (T cell enriched, immune infiltrated stroma, macrophage, and enriched) was prevalent significantly more often than in the maximum entropy null distribution in DII patients (Figure 5G, right, colored triangle). The CNM had highlighted that SC(1,3,4) was associated with inhibitions on the activation of T cells. In summary, in DII patients, a potentially inhibitory motif is actively formed, and in CLR patients, a potentially effector motif is actively formed.

Triangle (1,4,7) (T cell enriched, macrophage enriched, and vasculature) was actively assembled in CLR patients, but triangle (1,3,7) (T cell enriched, stroma enriched, and vasculature) was actively assembled in DII patients (Figure 5G, triangles in left/right). This can be interpreted as evidence that the cn1-cn7 (T cell enriched/vasculature) motif has an altered role in DII patients, by assembling with the stromal enriched CN. In addition, triangle (2,6,9) was actively assembled in DII patients, further corroborating that the signals generated in cn9 (granulocyte enriched), and their effects in opposition to those generated in the tumor (e.g. on DC phenotypes as suggested by the CNM) could be an ingredient of the antitumoral immune response in DII patients.

We observed that triangle (1,4,7), (T cell enriched, macrophage enriched, and vasculature), which was the most significantly enriched in the CLR patient group relative to the maximum-entropy null distribution, was present in 16/17 CLR patients and 17/18 DII patients. The fact that this motif was present in both groups but only actively assembled in CLR raised the question of whether it was an important component of the immune response. If so, there clinically significant differences in how this motif is used in each patient group would be expected. For each extension of triangle (1,4,7) by a single CN, we quantified the differences between patient groups in the frequency of patients with each such extension (Figure 5H). While many extensions were similarly represented across both patient groups, as would be expected (Figure 5H, motifs in middle of plot), the extension of this motif by the follicle CN 5 was enriched among CLR patients (Figure 5H, lower end of plot). On the other hand, there was a great enrichment of the motif formed by the extension of triangle (1,4,7) with cn2 (bulk tumor) adjacent to

Cell Systems
Article

CellPress

cn1 (T cell enriched) in DII patients (Figure 5H, upper end of plot), suggesting that the tumor incorporating itself into a motif could affect clinical outcomes.

We found that DII patients with at least one instance of the cn2-triangle(1,4,7) motif had dramatically worse survival outcomes than DII patients that did not have any instances of that motif (despite 17/18 DII patients having an instance of triangle (1,4,7)) (Cox Proportional Hazards regression p value = 0.006) (Figure 5I). Taken together, these data provide evidence that the (T cell-enriched, macrophage-enriched, and vasculature) motif is actively formed as part of an effective antitumoral response, and moreover, tumors can dynamically repurpose its assembly in a clinically significant manner. When the presence of the cn2-triangle(1,4,7) motif was included as a feature in combination with the functional state of CN-9 (i.e., the frequency of PD-1+ CD-4+ T cells in the granulocytic CN, which had previously been reported to positively associated with survival of DII patients (Schürch et al., 2020) in a Cox proportional hazards regression model on survival outcomes in DII patients, both features were highly significant (p < 0.005) (Figure S9). Thus, the biological programs mediating the negative effect on survival associated with the cn2-triangle(1,4,7) motif, and the positive effect of the functional state of cn9 are complementary.

We identified whether any motifs related to the triangle(1,4,7) motif were found only in patients without the cn2-triangle(1,4,7) motif, and observed that the only ones that were found in at least five patients with those CNs were complex ones (Figure 5J, visualized motifs and corresponding TMA cores), but these did not have significant association with survival. We observed cn9 in two of these motifs, indicating that simultaneously inhibiting the assembly of cn2-triangle(1,4,7) and inducing the adjacency and functional state of cn9 could synergize to help improve the outcomes of DII patients. This raised the question is the functional role of the cn2-triangle(1,4,7) motif emergent, or is it also obtained by any of its subcomponents? We assessed whether the presence of any submotifs of the cn2-triangle(1,4,7) motif were associated with survival (Figure 5K). In that graph the color intensity of the surrounding circle indicates the p value of the association with survival. All of the submotifs that were associated with survival required all four of the CNs 1,4,2, and 7, showing that assessment of the more complicated motif is required for clinical stratification of patients. Moreover, the graph also demonstrates how a common assembly (one with no stratification of patients by survival [Figure 5K, black arrows]) can be corrupted by the tumor to construct an iTME detrimental to patient health (Figure 5K, red arrows).

Through this approach for dissecting functional contributions to the CRC iTME, we have established the following: (1) the phenomenon of interactions between the local processes of CNs in SCs giving rise to emergent tissue signals is a common feature of tissue function across multiple tissue types and sizes, (2) complex motifs are not merely statistically emergent from tissue constraints on simpler motifs, but rather specific motifs are higher order (likely to be actively assembled), (3) such higher-order motifs can play a distinct, functional role from their simpler subcomponents, and (4) the distinct functional role of higher-order motifs can affect patient outcomes.

DISCUSSION

The TSs approach, which consists of identifying CNs, mapping their local interactions in SCs and the rules of their assembly into tissue motifs, provides a systematized strategy to infer (using high-parameter imaging data) how tissues achieve complex functions by composing simpler components.

The map of the SCs in the CRC iTME, as well as changes in expression of functional immune markers within them, suggests that cellular mechanisms of how the CRC iTME is wired to generate immune responses: (1) the local processes of the TLS and T cell-enriched CN are analogous to those of a lymphoid organ, and key features of the tissue schematic of HLT are implemented in TLS, (2) the local processes of the tumor generates antigen-dependent immune activation, (3) interactions between the local processes of the main tumor and tumor boundary interact to give rise to signals driving macrophage proliferation, (4) the local processes of the macrophage CN interact with those of the T-cell-enriched and immune-infiltrated stroma CNs leading to the generation of DC enhanced inhibition of T cells, and (5) the local processes of the compartmentalized region of the granulocyte enriched CN generate effects on DC phenotypes that oppose immune inhibitory local processes of the tumor. In addition, the way in which the known architectural optimizations of the germinal center response are reflected in the SCs formed by the CNs of HLT shows that CNMs can be an informative way to infer the nature of signal propagation between CNs.

The analysis of basic rules of HLT motif assembly aligned the three tissue types, identifying a core, conserved lymphoid motif, and its specialization to each tissue type. Furthermore, the basic rules showed that the architecture of the lymph node and tonsil were more similar than that of either to the spleen. Assessing the motif assembly of the CRC iTME, we identified evidence of conserved assembly and wiring between the TLS and the follicles of HLT. Furthermore, our statistical framework identified a higher-order tissue motif whose mutation was associated with survival outcomes for high-risk patients that was complementary to the one that had been previously identified. The tumor's corruption of the T-cell-enriched, macrophage-enriched, vasculature triangle motif may play a similar role in other iTMEs beyond that of CRC. Our results highlight that while the compositional nature of immune architecture enables different lymphoid tissues to specialize common immune programs, it also provides opportunities for the tumor to strategically insert itself into these programs.

Future work to define SCs that utilize different strategies from the CN combination one (for example those that directly model chemokine gradients or distance from vasculature) may elucidate more complex local processes and interactions. Extending the identification of motifs to those defined in terms of more complicated spatial ontologies (Bateman and Farrar, 2004) will increase the granularity with which assemblies of CNs and their functional role can be defined and understood. A limitation of our analysis is that tissue-volume sampling biases are not taken into account in the identification of rules. Therefore, statistical models for assessing the presence of ARs and changes within them will address such considerations of spatial sampling power. Extending ARs to incorporate more complicated qualitative spatial reasoning techniques (Dylla et al., 2017) and identifications of redundancies in such rules (Li et al., 2015) will enhance automated construction

Please cite this article in press as: Bhate et al., Tissue schematics map the specialization of immune tissue motifs and their appropriation by tumors, Cell Systems (2021), https://doi.org/10.1016/j.cels.2021.09.012

of TSs. Further, we propose here only that that understanding tissues compositionally (describing them with respect to compositions of interacting components, while maintaining registry of the functionality of such components) will be essential to dissecting order out of the apparent disorder of the tumor microenvironment. As such, other models for composition (Willems and Polderman, 1997; Fong et al., 2016, 2021; Courser, 2020) applied to tissues will enhance mapping efforts. Furthermore, constructing schematics of tissue in a compositional manner enables perturbations to complex tissue components to be designed in-silico from those to simpler ones, providing both a strategy for designing therapeutic strategies and validating the schematic.

Merging of our data and approach with ongoing tissue atlas projects and perturbational data should enable identification of associated genetic programs that are definitive for the structures and functions we have identified. To this end, experimental models, such as in-vitro-insert-based systems (Marshall, 2011), organoid (Rossi et al., 2018), and bioprinting (Murphy and Atala, 2014) platforms, could allow quantitative functional understanding of signals emergent from the signal propagation in CN assemblies. The local processes of CNs and their interactions in SCs could be quantitatively modeled through optimal transport (Schiebinger et al., 2019), pseudo-spatiotemporal analysis (Nowotschin et al., 2019), and field-based models (Opper and Saad, 2001) to provide insights into the emergent order present in the architecture of these tissues. Extending our sampling-based approach for identifying higher-order structures will enable establishment of the minimal relationships required to specify the assembly of tissues and will thus inform engineering of spatially organized systems such as those enabled by synthetic morphogen systems (Toda et al., 2020).

TSs constructed across other disease states could identify biomarkers and yield mechanistic insights into altered tissue behaviors associated with pathologies of the kind we have identified here. TSs could also be constructed across organismal development time from egg to fetus to adult stages to understand the developmental programs regulating CN assemblies and their behavior. Indeed, cross-species comparisons of TSs will reveal the architectural innovations enabling specialized and more effective tissue function. We expect that as data accumulate across multiple immune organs, perturbation conditions, age, and species comparisons that the "nuts and bolts" of the schematics of various tissue types will come together in a manner that allows its modulation toward correct function during disease and pathologies.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- ● KEY RESOURCES TABLE
- ● RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- ● METHOD DETAILS
  - ○ CODEX imaging data
  - ○ CN identification in the HLT dataset
  - ○ Barycentric coordinate projection

- ○ Identification of SCs for HLT
- ○ CN combination map for HLT
- ○ Instance segmentation for HLT dataset
- ○ Tissue graph for HLT dataset
- ○ Hashing of motifs, instances and extensions from tissue graphs
- ○ Identification of assembly rules
- ○ Identification of basic assembly rules
- ○ Identification of SCs for CRC dataset
- ○ Identification of differences in SC frequency between CLR and DII patients
- ○ Identification of cell types enriched in SC(1,2,4,6)
- ○ Instance segmentation for CRC dataset
- ○ Tissue graph for CRC dataset
- ○ Comparison of fragmentation between CLR/DII patients
- ○ Identification of differential extension frequencies between CLR and DII patients
- ○ Identification of higher-order 2-chains in CLR and DII patients
- ○ Construction of Level-1 Null Set
- ○ Sampling from the uniform distribution over the Level-1 Null Set
- ○ Hypothesis testing to identify higher-order triangles in CRC
- ○ Identifying extensions of tri(1,4,7)
- ○ Survival analysis
- ● QUANTIFICATION AND STATISTICAL ANALYSIS

Cell Systems
Article

CellPress

REFERENCES

Agasti, S.S., Wang, Y., Schueder, F., Sukumar, A., Jungmann, R., and Yin, P. (2017). DNA-barcoded labeling probes for highly multiplexed exchange-PAINT imaging. Chem. Sci. 8, 3080–3091. https://doi.org/10.1039/C6SC05420J.

Angelo, M., Bendall, S.C., Finck, R., Hale, M.B., Hitzman, C., Borowsky, A.D., Levenson, R.M., Lowe, J.B., Liu, S.D., Zhao, S., et al. (2014). Multiplexed ion beam imaging of human breast tumors. Nat. Med. 20, 436–442. https://doi.org/10.1038/nm.3488.

Arnol, D., Schapiro, D., Bodenmiller, B., Saez-Rodriguez, J., and Stegle, O. (2019). Modeling cell-cell interactions from spatial molecular data with spatial variance component analysis. Cell Rep 29, 202–211.e6.

Bateman, J., and Farrar, S. (2004). Towards a generic foundation for spatial ontology. In Proceedings of the third international conference on formal ontology in information systems, p. 248.

Binnewies, M., Roberts, E.W., Kersten, K., Chan, V., Fearon, D.F., Merad, M., Coussens, L.M., Gabrilovich, D.I., Ostrand-Rosenberg, S., Hedrick, C.C., et al. (2018). Understanding the tumor immune microenvironment (TIME) for effective therapy. Nat. Med. 24, 541–550.

Chib, S., and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. Am. Stat. 49, 327–335.

Courser, K. (2020). Open systems: a double categorical perspective. arXiv https://arxiv.org/abs/2008.02394.

Davidson-Pilon, C., Kalderstam, J., Jacobson, N., Reed, S., Kuhn, B., Zivich, P., Williamson, M., Abdeali, J.K., Datta, D., Fiore-Gartland, A., et al. (2021). CamDavidsonPilon/lifelines: v0.25.11 (Zenodo). https://doi.org/10.5281/zenodo.4683730.

Dylla, F., Lee, J.H., Mossakowski, T., Schneider, T., Delden, A.V., Ven, J.V.D., and Wolter, D. (2017). A survey of qualitative spatial and temporal calculi: algebraic and computational properties. ACM Comput. Surv. 50, 1–39. https://doi.org/10.1145/3038927.

Fong, B., Myers, D.J., and Spivak, D.I. (2021). Behavioral mereology: a modal logic for passing constraints. arXiv. https://doi.org/10.4204/EPTCS.333.19.

Fong, B., Sobociński, P., and Rapisarda, P. (2016). A categorical approach to open and interconnected dynamical systems. In Proceedings of the 31st annual ACM/IEEE symposium on Logic in Computer Science, pp. 495–504.

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, 968–981.e15. https://doi.org/10.1016/j.cell.2018.07.010.

Graham, D.M., and Appelman, H.D. (1990). Crohn's-like lymphoid reaction and colorectal carcinoma: a potential histologic prognosticator. Mod. Pathol. 3, 332–335.

Hardie, D.L., Johnson, G.D., Khan, M., and MacLennan, I.C. (1993). Quantitative analysis of molecules which distinguish functional compartments within germinal centers. Eur. J. Immunol. 23, 997–1004.

Jackson, H.W., Fischer, J.R., Zanotelli, V.R.T., Ali, H.R., Mechera, R., Soysal, S.D., Moch, H., Muenst, S., Varga, Z., Weber, W.P., and Bodenmiller, B. (2020). The single-cell pathology landscape of breast cancer. Nature 578, 615–620.

Kennedy-Darling, J., Bhate, S.S., Hickey, J.W., Black, S., Barlow, G.L., Vazquez, G., Venkataraaman, V.G., Samusik, N., Goltsev, Y., Schürch, C.M., and Nolan, G.P. (2021). Highly multiplexed tissue imaging using repeated oligonucleotide exchange reaction. Eur. J. Immunol. 51, 1262–1277.

Kepler, T.B., and Perelson, A.S. (1993). Cyclic re-entry of germinal center B cells and the efficiency of affinity maturation. Immunol. Today 14, 412–415.

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, 1373–1387.e19. https://doi.org/10.1016/j.cell.2018.08.039.

Li, S., Long, Z., Liu, W., Duckham, M., and Both, A. (2015). On redundant topological constraints. Artif. Intell. 225, 51–76.

Lin, J.R., Izar, B., Wang, S., Yapp, C., Mei, S., Shah, P.M., Santagata, S., and Sorger, P.K. (2018). Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes. Elife 7, e31657. https://doi.org/10.7554/eLife.31657.

Liu, Y.J., Barthélémy, C., de Bouteiller, O., Arpin, C., Durand, I., and Banchereau, J. (1995). Memory B cells from human tonsils colonize mucosal epithelium and directly present antigen to T cells by rapid up-regulation of B7-1 and B7-2. Immunity 2, 239–248.

Marshall, J. (2011). Transwell® invasion assays. In Cell Migration (Springer), pp. 97–110.

Murphy, S.V., and Atala, A. (2014). 3D bioprinting of tissues and organs. Nat. Biotechnol. 32, 773–785.

Nowotschin, S., Setty, M., Kuo, Y.Y., Liu, V., Garg, V., Sharma, R., Simon, C.S., Saiz, N., Gardner, R., Boutet, S.C., et al. (2019). The emergent landscape of the mouse gut endoderm at single-cell resolution. Nature 569, 361–367.

Nutt, S.L., and Tarlinton, D.M. (2011). Germinal center B and follicular helper T cells: siblings, cousins or just good friends? Nat. Immunol. 12, 472–477.

Opper, M., and Saad, D. (2001). Advanced Mean Field Methods: Theory and Practice (MIT Press).

Oyler-Yaniv, A., Oyler-Yaniv, J., Whitlock, B.M., Liu, Z., Germain, R.N., Huse, M., Altan-Bonnet, G., and Krichevsky, O. (2017). A tunable diffusion-consumption mechanism of cytokine propagation enables plasticity in cell-to-cell communication in the immune system. Immunity 46, 609–620.

Pipi, E., Nayar, S., Gardner, D.H., Colafrancesco, S., Smith, C., and Barone, F. (2018). Tertiary lymphoid structures: autoimmunity goes local. Front. Immunol. 9, 1952.

Rossi, G., Manfrin, A., and Lutolf, M.P. (2018). Progress and potential in organoid research. Nat. Rev. Genet. 19, 671–687.

Roy, V. (2020). Convergence diagnostics for Markov chain Monte Carlo. Annu. Rev. Stat. Appl. 7, 387–412.

Saka, S.K., Wang, Y., Kishi, J.Y., Zhu, A., Zeng, Y., Xie, W., Kirli, K., Yapp, C., Cicconet, M., Beliveau, B.J., et al. (2019). Immuno-SABER enables highly multiplexed and amplified protein imaging in tissues. Nat. Biotechnol. 37, 1080–1090. https://doi.org/10.1038/s41587-019-0207-y.

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

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. (2019). Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. Cell 176, 928–943.e22.

Schürch, C.M., Bhate, S.S., Barlow, G.L., Phillips, D.J., Noti, L., Zlobec, I., Chu, P., Black, S., Demeter, J., McIlwain, D.R., et al. (2020). Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front. Cell 182, 1341–1359.e19.

Ståhl, P.L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O., Huss, M., et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science *353*, 78–82.

Steiniger, B., Trabandt, M., and Barth, P.J. (2011). The follicular dendritic cell network in secondary follicles of human palatine tonsils and spleens. Histochem. Cell Biol. *135*, 327–336.

Steiniger, B.S., Raimer, L., Ecke, A., Stuck, B.A., and Cetin, Y. (2020). Plasma cells, plasmablasts, and AID+/CD30+ B lymphoblasts inside and outside germinal centres: details of the basal light zone and the outer zone in human palatine tonsils. Histochem. Cell Biol. *154*, 55–75.

Stoltzfus, C.R., Filipek, J., Gern, B.H., Olin, B.E., Leal, J.M., Wu, Y., Lyons-Cohen, M.R., Huang, J.Y., Paz-Stoltzfus, C.L., Plumlee, C.R., et al. (2020). CytoMAP: a spatial analysis toolbox reveals features of myeloid cell organization in lymphoid tissues. Cell Rep *31*, 107523.

Toda, S., McKeithan, W.L., Hakkinen, T.J., Lopez, P., Klein, O.D., and Lim, W.A. (2020). Engineering synthetic morphogen systems that can program multicellular patterning. Science *370*, 327–331.

Vickovic, S., Eraslan, G., Salmén, F., Klughammer, J., Stenbeck, L., Schapiro, D., Äijö, T., Bonneau, R., Bergenstråhle, L., Navarro, J.F., et al. (2019). High-definition spatial transcriptomics for in situ tissue profiling. Nat. Methods *16*, 987–990.

Wei, L., Chen, Z., Shi, L., Long, R., Anzalone, A.V., Zhang, L., Hu, F., Yuste, R., Cornish, V.W., and Min, W. (2017). Super-multiplex vibrational imaging. Nature *544*, 465–470. https://doi.org/10.1038/nature22051.

Willems, J.C., and Polderman, J.W. (1997). Introduction to Mathematical Systems Theory: A Behavioral Approach (Springer Science and Business Media).

Cell Systems
Article

 CellPress

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Spatial cellular data for HLT	Kennedy-Darling et al., 2021 (PI: Garry Nolan lab)	https://doi.org/10.1002/eji.202048891
Spatial cellular data for CRC and patient metadata	Schürch et al., 2020 (PI: Garry Nolan)	https://github.com/nolanlab/TissueSchematics https://doi.org/10.1016/j.cell.2020.07.005
Software and algorithms		
TissueSchematics jupyter notebooks	This paper	https://doi.org/10.5281/zenodo.5494009

RESOURCE AVAILABILITY

Lead contact
Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Garry Nolan (gnolan@stanford.edu).

Materials availability
This study did not generate or utilize any reagents.

Data and code availability
- The HLT and CRC CODEX data and associated metadata utilized are published in (Kennedy-Darling et al., 2021) and (Schürch et al., 2020) respectively.
- All code has been deposited at www.github.com/nolanlab/TissueSchematics/. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

CODEX imaging data
Raw imaging data, segmented single-cells and cell-type clusters types for the HLT dataset were obtained from (Kennedy-Darling et al., 2021). Segmented single-cells, cell-type clusters, and CN assignments for the CRC dataset were obtained from (Schürch et al., 2020).

CN identification in the HLT dataset
CNs in the HLT dataset were identified by clustering cells based on the cell-type composition of their 20 nearest neighboring cells. For each cell, a window of size 20 nearest neighboring cells (including the center cell) as measured by Euclidean distance between X/Y coordinates was obtained. These windows were then clustered by the composition of their microenvironment with respect to the 25 previously identified cell types. Specifically, each window was converted to a vector of length 25 containing the frequency of each of the 25 cell types amongst the 20 neighbors. We then clustered these windows using Python's scikit-learn implementation of MiniBatchKMeans with $k = 11$. Each cell was then allocated to the same CN as the window in which it was centered. To validate the CN assignment, these allocations were overlaid on the original tissue images and cross compared with expression of the characteristic markers of the cell types enriched in that CN. We observed that the original follicular CNs did not accurately reflect the distribution of marker expression that could be observed in the original data, because the cell types in the original publication merged phenotypic subsets of B-cells. We finely re-clustering all cells across the tissues using the K-means algorithm with a $k=100$, identifying 11 clusters that were abundant in the follicle, and re-clustering the original follicular CNs into three new CNs using these cell types.

Barycentric coordinate projection
Windows of 100 nearest neighbors were extracted around each cell, and frequencies of CN assignments to each individual cell in each window were computed. Those windows where cnB, cnLZ, and cnDZ comprised more than 90% of the cells were selected and projected linearly under the linear map sending the unit coordinate vectors (i.e., 100% of cnB, cnLZ and cnDZ) to a point of an equilateral triangle. The projected points were colored by the CN assignment at the center of the window.

Identification of SCs for HLT

Windows of 100 nearest neighbors were extracted around each cell, and frequencies of CN assignments in each window to each individual cell were computed. Windows were assigned to the minimal combination of CNs such that 90% of the cells in a window were assigned to a CN from that combination. Figure S5 shows that neither which combinations, or their relative abundances are sensitive to small changes in the choice of 100 for window size or 90% for the minimal parameter.

CN combination map for HLT

A directed graph was built from the abundant combinations (those more than .001% of total cells) using the Python networkx library, where nodes were combinations and edges were added between a source combination and a target combination if the combination at the target could be obtained by adding one CN to the combination at the source. These were visualized using the Python matplotlib library.

Instance segmentation for HLT dataset

For each CN and tissue, a binary image was created from the x, y coordinates of each cell; pixels containing cell centers assigned to the CN of interest were assigned a value of 1. The size of these images was such that the x, y coordinates of the original cell were downsampled by a factor of 50. For CNs BT, PI LZ, DZ, binary images were dilated using the SciPy ndi binary dilation function. The connected components of the binary images were identified using the SciPy ndi label function with default parameters. Connected components that were less than 20 pixels in size were assigned to background. The collection of instances of each CN were the non-background-connected components in each tissue.

Tissue graph for HLT dataset

A graph for each tissue was built with nodes corresponding to instances of each CN, and edges corresponding to pairs of instances that spatially intersected (i.e., that had at least 1 pixel in common) in the binary images. A combined graph for the tonsil was constructed as the union of the graphs from each tonsils. Graphs were represented using the python networkx library.

Hashing of motifs, instances and extensions from tissue graphs

Only one CN name was only allowed to appear in the motifs we considered. As a result, we could represent motifs programmatically (we used python) as a pair (A,B), where:

- A was a lexicographically ordered tuple of lexicographically ordered pairs of CN names, representing edges in the graph of a motif
- B was a lexicographically ordered tuple of CN names, representing the vertices of a motif.

A motif instance – a collection of CN instances respecting the adjacency structure specified by the motif was programmatically represented as a lexicographically ordered tuple of CN instances.

Using this convention, we constructed a python dictionary (a map of key-> value pairs) whose keys were motifs, and values were sets of motif instances (termed the motif-instances map). Specifically, we initialized the algorithm by adding 2-chains and their corresponding instances by iterating over the edges of the tissue graph to the dictionary. Next, we iteratively extended the motifs already in the keys of the dictionary by iterating over their instances and adding adjacent vertices of the graph, adding these instances to the set of instances for the correspondingly extended motifs and therefore adding to the motif-instance map. Since we were iterating over each motif instance to identify its extensions, we simultaneously constructed another dictionary, termed the extension-in-stances map, whose keys were pairs (m1,m2) where m1 is a submotif of m2 and values corresponded to sets of instances of m1 that could be extended to m2, and a directed graph with edges (m1,m2), termed the extension graph.

We only extended motifs with at most 5 CN names, so the motifs we assessed had at most six CN names.

Identification of assembly rules

The assembly rules were identified by selecting edges in the extension graph with source motifs with at least 5 instances for which at least 70% of the instances of the source motif could be extended to the target motif. We validated that this 70% threshold did not affect the rules identified, capturing the major structure present in extensions of HLT instances to 2 chains (Figure S7). This collection of edges was further pruned by enforcing transitivity; so that rules were only selected if they could be composed (i.e., that for every submotif of a given motif, at least 70% of the instances of that submotif could be extended to the larger motif of the rule). This subgraph of the extension graph was termed the rule graph.

Identification of basic assembly rules

The basic rule graph was obtained from the rule graph as follows. First, each edge in the extension graph was assigned a 'difference', corresponding to the specific edge between CNs added to the motif of the source to obtain the motif at the target. For example, if an edge in the rule graph corresponded to the motif (A-B) being extended to the motif (A-B-C), the difference assigned to that edge of the rule graph would be (B-C). An edge was added to the basic rule graph if the difference assigned to it was not equal to the difference assigned to any of its ancestors' outgoing edges.

Cell Systems
Article

CellPress

Identification of SCs for CRC dataset

Cores that had 750 cells or fewer were excluded (corresponding to 13/140 cores). SCs were identified similarly to the HLT dataset: windows of 50-nearest neighbors were extracted around each cell, and frequencies of CN assignments in each window to each individual cell were computed. Windows were assigned to the minimal combination of CNs such that 90% of the cells in a window were assigned to a CN from that combination. We then identified abundant SCs as those with at least 25 cells in each of at least 10 patients. This resulted in in 55 abundant SCs.

Identification of differences in SC frequency between CLR and DII patients

The frequency of each abundant SC was computed for each patient. Mann-Whitney U tests were performed using the python function scipy.stats.mannwhitneyu to obtain a p-value for differences between these groups. FDRs were computed using the Benjamini-Hochberg procedure as implemented in the python function statsmodels.stats.multitest.fdrcorrection. Reported differences are at a FDR of 0.05.

Identification of cell types enriched in SC(1,2,4,6)

The following expression of functional markers was obtained from the manually gated data reported in (Schürch et al., 2020). CD4+ T cells (PD-1, Ki67), CD8+ T cells (PD-1, Ki67), Tregs (PD-1, Ki67), B cells (PD-1, Ki67), CD68+CD163+ macrophages (PD-1, Ki67, VISTA), CD11c+ DCs (PD-1, Ki67, VISTA). Next, for each SC and cell subset, a null distribution for the percentage of cells expressing a given functional marker was generated as follows. For each SC and cell subset and TMA core, 30000 samples of that cell subset (ignoring expression of the functional marker) of the same size as the number of the given cell subset in the given SC in each TMA core were obtained. These cells were pooled to generate a sample of cells across the tissues consisting of the same number of the immune cell subset as the number of that immune cell subset in the given SC across the tissues. Next, the percentage of the sampled subset of cells that were positive for each functional marker was computed in each sampled subset, thereby generating a null distribution for marker in the SC. The observed percentage of the immune cells in the SC was compared to this null distribution. A p-value was obtained as the minimum of:

- 1- Average number of null samples whose frequency of functional marker expression exceeded the frequency of functional marker expression in the immune cell subset in the SC
- 1- Average number of null samples whose frequency of functional marker expression was less than the frequency of functional marker expression in the immune cell subset in the SC

FDRs were obtained using the python statsmodels.stats.multitest.fdrcorrection function, and the reported results are at a FDR of 0.05.

Instance segmentation for CRC dataset

Since each TMA core was small, a different approach was used to segment instances. A 10-nearest neighbor graph for the cells of each TMA core was constructed using the sklearn.nearest_neighbors module. The connected components of this graph were identified using the scipy.sparse.csgraph module. Instances which consisted of fewer than 5 cells were discarded

Tissue graph for CRC dataset

A graph for each TMA core was formed whose edges were the segmented CN instances from that core and whose edges were CN instances in each TMA core that were adjacent. Two instances were considered adjacent if one contained a cell that neighbored another in the other.

Comparison of fragmentation between CLR/DII patients

The number of nodes and average degree for the tissue graph of each TMA core was computed, which is what is displayed in Figure 4F. For each patient, the average of these was taken across the tissue graphs of their 4 TMA cores. The p-values for differences between groups were computed using the python scipy.stats implementation of the Mann Whitney U-test.

Identification of differential extension frequencies between CLR and DII patients

A null distribution for the difference in extension frequencies was generated by permuting patient assignments to CLR or DII. Since each patient had multiple TMA cores, permutations had to be conducted at the patient level and not at the TMA core level. Specifically, for each permuted patient group assignment, two graphs were obtained by taking the union of the tissue graphs for each TMA core assigned to patients from the respective permuted group. In each of these graphs, the instances of each CN were enumerated, and the frequency that each of those instances was extended to a 2-chain was computed. For each possible extension, the difference between the frequency of extension in the permuted assignment to CLR and in the permuted assignment to DII was obtained. A one-sided hypothesis test was conducted for each extension against the null hypothesis that the difference in extension frequency was 0. The rejection p-value was the frequency that the permuted difference exceeded the observed difference when positive, or vice versa when negative. 10000 permutations were conducted. p-values were Bonferroni corrected by multiplying by twice the number of tests conducted.

CellPress

Cell Systems
Article

Identification of higher-order 2-chains in CLR and DII patients

A tissue graph for each patient group was obtained by taking the union of the tissue graphs of each TMA core. We considered the Level-0 null-set defined as follows. The level 0 null-set was defined as graph the set of CN-assignments obtainable by a sequence of valid transpositions of CN-assignments, where a valid transposition was:

- Any transposition of the CN assignment of two nodes in same TMA core.

The Level-0 null-set can be seen to be equivalent to the set of CN assignments obtainable by arbitrary product of permutations of the assignments within each core, as well as the set of CN assignments that fix the number of vertices in each core assigned to each CN. The maximum entropy null distribution – given by the uniform distribution on this set – could therefore by sampled directly using random permutations from the numpy.random library.

Only those 2-chains with at least 5 instances were considered as candidate higher-order structures. One-sided hypothesis tests were conducted for those 2-chains against the hypothesis that the observed count arose as the count of that 2-chain as a sample assignment from the null distribution. The null hypothesis rejection p-value was estimated as the smaller of:

- 1-(the proportion of permutations for which the observed count was greater than the count in the permuted assignment)
- 1- (the proportion of permutations for which the observed count was less than the count in the permuted assignment).

To identify significant 2-chains, p-values were Bonferroni corrected by multiplying by twice the number of tests conducted in each patient group.

Construction of Level-1 Null Set

The Level-1 Null Set for the tissue graphs of each group was defined as the set of colorings obtainable by a sequence of valid transpositions of color-assignments. We considered two distinct notions of valid transposition.

- The 'exact' level-1 null set was obtained by considering a valid transposition to be one between two nodes from the same TMA core whose neighboring nodes had the exact same collection of colors taking into account multiplicity (a given CN instance could be adjacent to multiple instances of the same CN).
- The 'inexact' level-1 null set was obtained by considering a valid transposition to be one between two nodes from the same TMA core whose neighboring nodes had the same set of colors, counting each color at most once.

From these definitions, we can see the following:

- The exact level-1 null set is a subset of the inexact level-1 null set.
- The exact level-1 null set exactly preserves the number of instances of each 2-chain, but the inexact level-1 null set does not (since a valid transposition in the inexact case could for example swap the colors of two nodes that have the same set but different numbers of adjacent colors).
- The exact level-1 null set is a subset of but not in general equal to the set of permutations that fixes the number of each 2-chain. This is because some permutations of the original coloring that fix the number of 2-chains in a colored graph might require permuting the color assignments of more than two nodes simultaneously. For example, considering a triangle with 3 different colors to each node, there are no valid transpositions since each node has a different set of neighboring colors, but any permutation of the coloring fixes the number of 2-chains.

In either of these cases, it is not possible to directly sample a random coloring from the Level-1 null set because of the constraints defining valid transpositions.

Sampling from the uniform distribution over the Level-1 Null Set

We implemented the Metropolis-Hastings (MH) algorithm to generate samples of colorings from the uniform distribution over the exact and inexact Level 1 null-sets. MH is a Markov Chain Monte Carlo algorithm. This algorithm operates by generating candidate transpositions with a certain probability for the tissue graph of each TMA core, and accepting them with a certain probability termed the acceptance probability, and otherwise rejecting them. The candidate transpositions and acceptance probability are chosen, such that the distribution of generated samples after hundreds of thousands of iterations is approximately the uniform distribution on the level 1 null set. To describe our algorithm, we will first give a brief overview, then discuss the implementation, and subsequently the steps we took to validate it.

We first give a brief overview of the approach. Starting with a coloring x_t in the level 1 null set, we generate a candidate new coloring x' by applying a transposition chosen with probability $g(x'|x_t)$. Crucially, this candidate transposition, by how it is selected, also belongs to the level 1 null set. Since the target distribution is uniform on the level 1 null-set, and our sampled transposition belongs to this set, the Metropolis acceptance probability has a simple formula:

$$A = min(1, g(x|x') / g(x'|x))$$

as described in(Chib and Greenberg, 1995). We obtain a sample from a random uniform variable in [0,1], r, and set $x_{t+1} = x'$ if $r < A$, or set $x_{t+1} = x_t$ if $r > A$.

Cell Systems
Article

⟳ CellPress

To generate the candidate transposition, we construct a map assigning vertices its equivalence class, which is defined as either:

- The set of colors that its neighboring vertices have, in the case of the inexact Level-1 null set
- The exact collection of colors that its neighboring vertices have taking into account multiplicity.

We simultaneously construct a map assigning an equivalence class to the set of vertices that belong to that equivalence class. Using these maps, we identify for each vertex, the other vertices in the graph that have distinct CN assignments but the same equivalence class, i.e. we identify candidate transpositions that are allowed and alter the overall coloring. We select a vertex that has a candidate transposition involving it uniformly, and also choose a vertex with the same equivalence class that has a different color to it uniformly.

This means that our chance of selecting a candidate transposition of colors between vertices s and t is given by:

$$\left(\frac{1}{number\ of\ vertices\ part\ of\ a\ candidate\ transposition}\right)$$
$$\times \left(\frac{1}{number\ of\ vertices\ in\ same\ equivalence\ class\ as\ vertex\ s\ with\ different\ color}\right.$$
$$\left.+\frac{1}{number\ of\ vertices\ in\ same\ equivalence\ class\ as\ vertex\ t\ with\ different\ color}\right)$$

Next, since the equivalence classes change only for the vertices adjacent to the transposed ones, we update the maps assigning vertices to their equivalence class, and equivalence classes to their set of vertices. We compute $g(x'|x)$ by the same formula as above, but with the updated maps. This enables computation of $g(x|x')$ as with $g(x'|x)$, which enables computation of the MH acceptance probability $A = \min(1, g(x|x')/g(x'|x))$. A random uniform sample r is drawn from [0,1]. If $r<A$, the new state x' is accepted and taken to the next iteration. If $r>A$, the transposition is rejected, and x and the previous maps are taken forward to the next iteration. The rejection is implemented by applying the transposition again and updating the maps, since this is more efficient than copying the entire map.

Since in each step, the coloring differs by at most 1, the samples are autocorrelated, but we require independent samples. This is a general challenge for MCMC algorithms, and how well the samples match the target distribution is typically validated by inspecting the traces and autocorrelation functions of samples over sequential MH steps to visually assess convergence and exploration of the state space(Roy, 2020). We generated 25000 samples for each TMA core each 8000 Metropolis-Hastings steps apart (2E8 samples), discarding the first 1E5 samples, using both the exact and inexact level 1 null set. This thinning of samples was performed for memory purposes. We computed the counts of each triangle observed in the original tissue graph for each selected core as well as each MCMC sample. We then visualized traces for enrichment (log_2 counts of the motif in the observed coloring/counts in the sample) across successive samples (Figure S10A, traces in either for three spots column shown from blue to red). This analysis revealed three kinds of TMA core. First, there were cores in which the enrichment in the observed graph relative to the inexact Level-1 null set and the exact Level-1 null set agreed (Figure S10A, top row traces). In these cases, the autocorrelation for each motif in the enrichment was zero for all lags (Figure S10A, top row, autocorrelation plots). In the second case, cores had many motifs that were enriched or depleted relative to the inexact level-1 null set, but considerably fewer relative to the exact level-1 null set (Figure S10, second row), but autocorrelations in both cases were essentially zero for all lags and motifs. Finally, there were a few TMA cores for which every motif was highly deviant from the samples from the maximum entropy distribution on the inexact level-1 null set, but considerably fewer motifs were deviant relative to the exact level-1 null set (Figure S10A, third row). In the third case, there were autocorrelations for some motifs with respect to the inexact level-1 null for many lags, but these also went to zero for lags greater than 50. We therefore visualized each spot by the distribution of motifs with respect to their autocorrelation at different lags, considering samples from the exact and inexact level-1 null sets (Figures S10B and S10C respectively). This showed quantitively that the autocorrelations for lags beyond 100 were essentially negligible for both the inexact and exact level-1 maximum entropy distributions, and thereby that our collection of 25000 MCMC samples every 8000 steps was a reasonable one, with which to assess motifs that deviated.

To generate sampled tissue graphs for each patient group, we uniformly sampled 25000 of the colorings generated by MH for each TMA core from that patient group, and pooled the counts of triangles from each core resulting in 25000 samples.

Our implementation was in C++, interfacing with python using the pybind11 library. This achieved a 100x speedup over an initial python implementation.

Hypothesis testing to identify higher-order triangles in CRC

Triangles that were present in at least 10 patients were considered in each patient group. Hypothesis testing was conducted as with 2-chains. The null hypothesis rejection p-value was estimated as the smaller of:

- 1-(the proportion of sample colorings for which the observed count was greater than the count in the count in the sample coloring)
- 1- (the proportion of sample colorings, for which the observed count was less than the count in the sample coloring).

To identify significant triangles, p-values were combined across patient groups and subsequently FDR adjusted using the Benjamini-Hochberg procedure from the python function statsmodel.stats.fdrcorrection, prior to identifying those motifs significant at FDR = 0.05.

We found no significant motifs in either patient group when using samples from the maximum entropy distribution over exact Level-1 null set (Figure S10D, both volcano plots). In addition, deviation from this distribution was generally very small (Figure S10D, positions along x axis). Thus, the uniform distribution over the exact level-1 null set, while easier to sample from, was too conservative a null distribution relative to which motifs could be identified. As such, the significant results in Figure 5G are those obtained with respect to the maximum entropy distribution over the inexact level-1 null set.

Identifying extensions of tri(1,4,7)

The motifs, instances and extensions for the tissue graphs of each TMA core were hashed as described above. In addition, a map was computed assigning each motif to the set of patients with at least one instance of that motif. Thus, we found the outgoing edges of tri(1,4,7) in the extension graph, and identified the number of patients in each patient group with that extension.

Survival analysis

We used the survival data from the previous paper to assess survival in DII patients only. This had overall survival times for 18 total patients with 5 censored observations. The lifelines (Davidson-Pilon et al., 2021) python package was used to estimate a Cox Proportional-Hazards regression model with a binary feature for whether or not a given patient had an instance of the cn2-tri(1,4,7) motif. The lifelines KMFitter object was used to produce the Kaplan-Meier curve.

A second Cox Proportional Hazards regression model was fit by including as a feature the the frequency of PD-1+ CD4+ T cells in cn9 alongside the presence of at least one instance of the cn2-tri(1,4,7) motif.

For the graph associating survival with submotifs, we identified submotifs of the cn2-tri(1,4,7) motif as ancestors in the extension graph. We assessed association for the presence of at least one instance of each submotif with survival in DII patients as above.

QUANTIFICATION AND STATISTICAL ANALYSIS

Details of quantification and statistical tests are described in method details and legends. All results were corrected for multiple hypothesis testing as described there.