ATHENA: Analysis of Tumor Heterogeneity from Spatial Omics Measurements

Supplementary Material

Adriano Martinelli[1,2], Maria Anna Rapsomaniki[1,*]

[1] IBM Research Europe, Zurich, Rüschlikon, Switzerland

[2] ETH Zurich, Switzerland

* aap@zurich.ibm.com

1 ATHENA overview

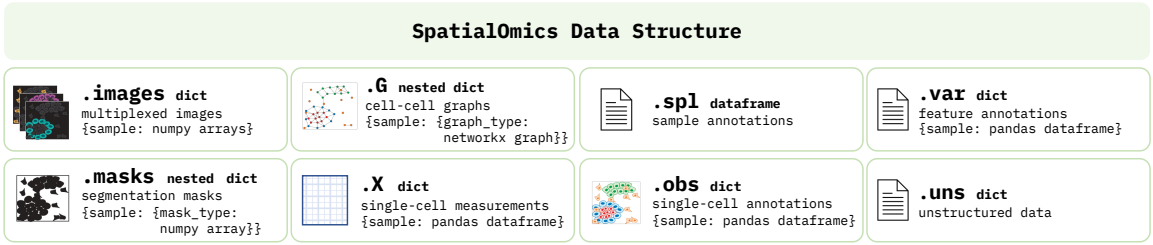
We present ATHENA (Analysis of Tumor HEterogeNeity from spAtial omics measurements), an open-source computational framework that brings together established and novel scores able to capture the heterogeneity of the tumor ecosystem. ATHENA supports any spatial omic data modality (e.g., imaging mass cytometry (IMC), multiplexed ion beam imaging (MIBI), seqFISH, Visium), as well as standard tissue imaging data (e.g., multiplexed immunohistochemisty (mIHC) or immunofluorescence (mIF). Those technologies profile samples at different resolutions (e.g., IMC or MIBI produce data at subcellular resolution, and Visium sequences multiple cells per spot). For simplicity, throughout this document and figures we use *cell* as the fundamental entity of an individual observation, but we note that in practice the exact nature of that observation depends on the specifics of the data acquisition technology and the preprocessing of the acquired data.

ATHENA is implemented in a highly modular fashion, based on two main components:

- SpatialOmics, a new data structure inspired by AnnData[1].
- SpatialHeterogeneity, a module that enables the computation of various heterogeneity scores.

1.1 SpatialOmics Data Structure

The SpatialOmics class is designed to accommodate storing and processing spatial omics datasets in a technology-agnostic and memory-efficient way. A SpatialOmics instance incorporates multiple attributes that bundle together the multiplexed raw images with the segmentation masks, cell-cell graphs, single-cell values, and sample-, feature- and cell-level annotations, as outlined in Supplementary Figure 1 below. Since ATHENA works with multiplexed images, memory complexity is a problem. SpatialOmics stores data in a HDF5 file and lazily loads the required images on the fly to keep the memory consumption low. The SpatialOmics structure is sample-centric, i.e., all samples from a spatial omics experiment are stored separately by heavily using Python dictionaries.



Supplementary Figure 1: **Overview of SpatialOmics data structure**. Boxes indicate different attributes, their content, Python data type and dimensions.

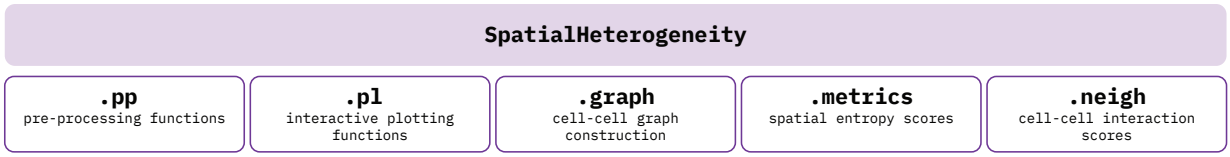
Specifically, each SpatialOmics instance contains the following attributes (Figure 1):

- .images: A Python dictionary (length: #samples) of raw multiplexed images, where each sample is mapped to a numpy array of shape: #features x image_width x image_height.
- .masks: A nested Python dictionary (length: #samples) supporting different types of segmentation masks (e.g., cell and tissue masks), where each sample is mapped to an inner dictionary (length: #mask_types), and each value of the inner dictionary is a discrete numpy array of shape: image_width x image_height.
- .G: A nested Python dictionary (length: #samples) supporting different topologies of graphs (e.g., knn, contact or radius graph), where each sample is mapped to an inner dictionary (length: #graph_types), and each value of the inner dictionary is a networkx [2] graph.

- .X: A Python dictionary of single-cell measurements (length: #samples), where each sample is mapped to a pandas dataframe of shape: #single_cells x #features. The values in .X can either be uploaded or directly computed from .images and .masks.
- .spl: A pandas dataframe containing sample-level annotations (e.g., patient clinical data) of shape: #samples x #annotations.
- .obs: A Python dictionary (length: #samples) containing single-cell-level annotations (e.g., cluster id, cell type, morphological features), where each sample is mapped to a pandas dataframe of shape: #single_cells x #annotations.
- .var: A Python dictionary (length: #samples) containing feature-level annotations (e.g. name of protein/transcript), where each sample is mapped to a pandas dataframe of shape: #features x #annotations.
- .uns: A Python dictionary containing unstructured data, e.g. various colormaps, experiment properties etc.

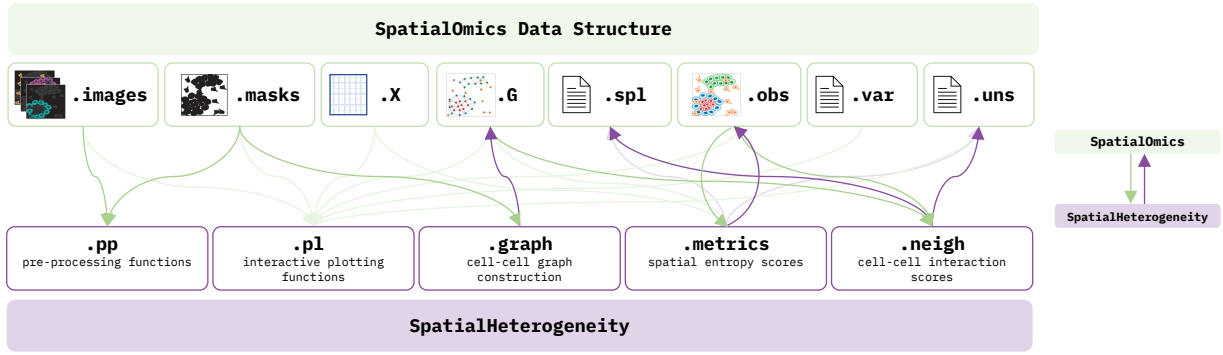
1.2 SpatialHeterogeneity module

The SpatialHeterogeneity module is the heart of ATHENA that implements all visualization, processing and analysis steps integral to its functionalities. SpatialHeterogeneity consists of the following 5 submodules, each one performing different tasks as outlined in Supplementary Figure 2 below:



Supplementary Figure 2: **Overview of SpatialHeterogeneity module**. Boxes indicate different submodules performing different functions.

SpatialHeterogeneity is tightly interwoven with SpatialOmics (see Supplementary Figure 3), in the sense that the submodules of SpatialHeterogeneity take as input various aspects of the data as stored in SpatialOmics (green arrows) and, at the same time, stores computed outputs back into different attributes of SpatialOmics (purple arrows).



Supplementary Figure 3: **Information flow between SpatialOmics and SpatialHeterogeneity.** Arrows indicate flow of information from SpatialOmics to SpatialHeterogeneity (light green) and vice versa (light purple).

- .pp works with .images and .masks and facilitates image pre-processing functions, such as extraction of cell centroids. ATHENA requires segmentation masks to be provided by the user.
- .pl supports plotting all aspects of the data, including raw images, masks, graphs and visualizes different annotations as well as results of computed heterogeneity scores. The plots can be either static or interactive, by exploiting the Python image viewer napari.
- .graph constructs cell-cell graphs from the cell masks using three different graph builders (kNN, radius and contact, details in Section 1.3). The resulting graphs are saved back to the .G attribute of SpatialOmics.
- .metrics uses the cell-cell graphs, the single-cell values (in .X) and cell annotations (in .obs) to compute a number of diversity scores, including sample richness (number of distinct cell subpopulations/clusters/clones) and abundance (relative proportions of species), and information-theoretic scores, (namely Shannon, Simpson, quadratic, or Renyi entropy, Hill numbers), either at a global, sample level (saved in .spl), or at a local, single-cell-level (saved in .obs) that incorporates the spatial information.
- .neigh implements a number of neighborhood or spatial statistics methods, namely infiltration score, Ripley's K and neighborhood analysis scores [3]. Results are saved in .spl and .uns.

1.3 Graph construction

The .graph submodule of SpatialHeterogeneity constructs a graph representation of the tissue using the cell masks extracted from the high-dimensional images. The graph construction module implements three different graph flavors that capture different kinds of cell-cell communication:

1.3.1 Contact graph

The *contact graph* representation mimics juxtacrine signaling, where cells exchange information via membrane receptors, junctions or extracellular matrix glycoproteins. The contact graph is constructed by binary dilation of the single-cell masks and then connecting cells that overlap after dilation. Dilation is a morphological modification of the original image, by which a structuring element S (a mask of arbitrary shape) is applied to the image [4]. If the structuring element placed at coordinate $\mathbf{x}$ and the cell masks in the image overlap, pixel $\mathbf{x}$ is set on, leading to an enlargement of the original cell mask. Other cell masks that intersect with the enlarged cell mask are considered neighbors in the contact graph.

1.3.2 Radius graph

The *radius graph* representation mimics paracrine signaling, where signaling molecules that are secreted into the extracellular environment interact with membrane receptors of neighboring cells and induce changes in their cellular state. The communication distance of paracrine signaling is determined by diffusion and is thus limited to relatively short distances. ATHENA's radius graph module builds on top of the scikit-learn implementation. For each cell, the centroid of the cell mask is used as its coordinate. The value of the radius can be adapted depending on the underlying data, with a default value of 36 pixels.

1.3.3 kNN graph

The *k-nearest neighbor (k-NN) graph* representation was chosen because of its popularity and application in digital pathology [5]. From a biological perspective, this representation is the most arbitrary, since it is not restricted by spatial distances or local cell densities, and leads to representations in which all cells have a minimum number of k neighbors. Similarly to the radius graph module, the kNN graph module builds on top of the scikit-learn implementation, and, for each cell, the centroid of the cell mask is used as its coordinate. The value of k is tunable, with a default value of $k = 5$, based on related work [5]. We note that a neighboring relationship is not a symmetric relationship (i.e., for any two cells i, j, the fact that cell i is a neighbor of cell j does not necessarily imply that cell j is a neighbor of cell i).

1.4 Heterogeneity quantification

ATHENA brings together a number of established as well as novel scores that enable the quantification of tumor heterogeneity in a spatially-aware manner, borrowing ideas from ecology, information theory, spatial statistics, and network analysis. As previously discussed, the majority of the scores included in ATHENA require as input a graph representation of the data and a single-cell-level phenotypic annotation. The latter can be achieved either via gating/manual thresholding, or by clustering of the single-cell profiles and using the cluster labels as phenotypes. Depending on the underlying mathematical foundations, the heterogeneity scores included in ATHENA can be classified into the following categories: (i) *spatial statistics* scores that quantify the degree of clustering or dispersion of each phenotype individually, (ii) *graph-theoretic* scores that examine the topology of the tumor graph, (iii) *information-theoretic* scores that quantify how diverse the tumor is with respect to different phenotypes present and their relative proportions, and (iv) *cell interaction* scores that assess the pairwise connections between different phenotypes in the tumor ecosystem. A summary of the implemented scores can be found in Supplementary Table 1; we next provide explanations and detailed mathematical definitions of these scores.

1.4.1 Spatial statistics scores

Spatial statistics scores quantify the localization of observations in space (*events*). **Ripley's** K is a commonly employed spatial statistic score that determines if events deviate from spatial homogeneity [6]. In the context of tumor heterogeneity, Ripley's K has been previously used to analyze the spatial heterogeneity of different types of immune cells using multiplexed IHC images [7, 8]. The K function is defined as:

$$K(r) = \lambda^{-1}\mathbf{E}\left[\text{number of events within distance } r \text{ of a randomly chosen event}\right] \tag{1}$$

where λ is the density of events. Ripley's K characterises the observed events at various distances r. To compute the expected number of events, one needs to assume a process that generates events in space, with the simplest and most broadly used model being the Poisson process that has a closed-form solution:

$$K(r) = \pi r^2 \tag{2}$$

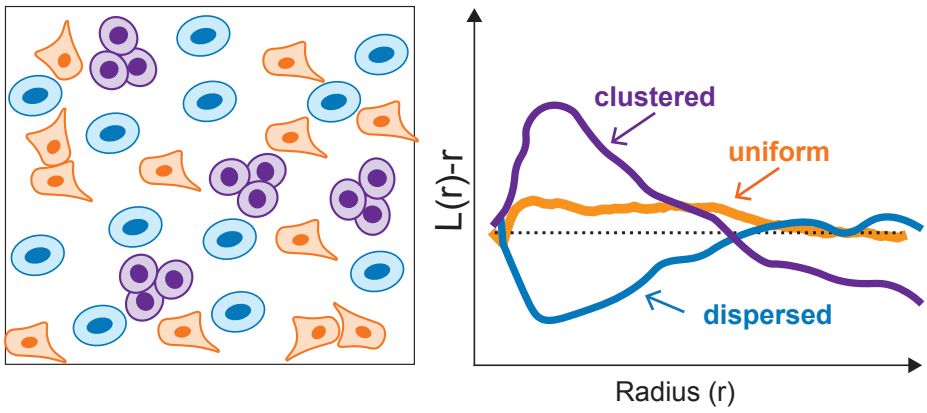
Estimating $\hat{K}(r)$ from a data set with N observations collected in an area A is performed by first estimating the density of the sample as $\hat{\lambda} = \frac{N}{A}$. Ignoring edge effects, the numerator can be estimated by $N^{-1}\sum_i \sum_{j\neq i} I(d_{ij} < r)$, where d_{ij} is the distance between two observations i and j and $I(x)$ is the indicator function that is equal to 1 if condition x is fulfilled, and 0 otherwise. However, this estimate ignores the fact that the study area is chosen arbitrarily and observations that are located outside of this area are not counted, which is especially problematic at large values of r. Various edge corrections have been proposed to counteract this bias. Ripley himself suggested to use a weight function $w(x_i, x_j)$ [9] that has a maximum value of 1 if a circle centered at x_i and passing through x_j is enclosed in the study area. On the other hand, if part of the circle is outside of the study area, then $w(x_i, x_j)$ is the proportion of the circumference of the circle that falls in the study area:

$$\hat{K}(r) = \lambda^{-1}\sum_i \sum_{i\neq j} w(x_i, x_j)^{-1}\frac{I(d_{ij} < r)}{N} \tag{3}$$

$\hat{K}(r)$ is commonly estimated for $r < \min(l_x, l_y)/2$, where l_x and l_y are the dimensions of the study area. Given $\hat{K}(r)$, one can test if the observed events correspond to a homogeneous Poisson process, i.e., if $\hat{K}(r) = \pi r^2$ for all r. In practice, the variance-stabilizing transform $\hat{L}(r) = \left[\hat{K}(r)/\pi\right]^{1/2}$ is commonly used. Spatial homogeneity is assessed by observing the plot of $\hat{L}(r) - r$ that should be equal to zero for uniformly distributed events in space (see Supplementary Figure 4 for an example). ATHENA supports the computation of Ripley's K and the variance-stabilizing transform $L(r)$ in the .neigh submodule.

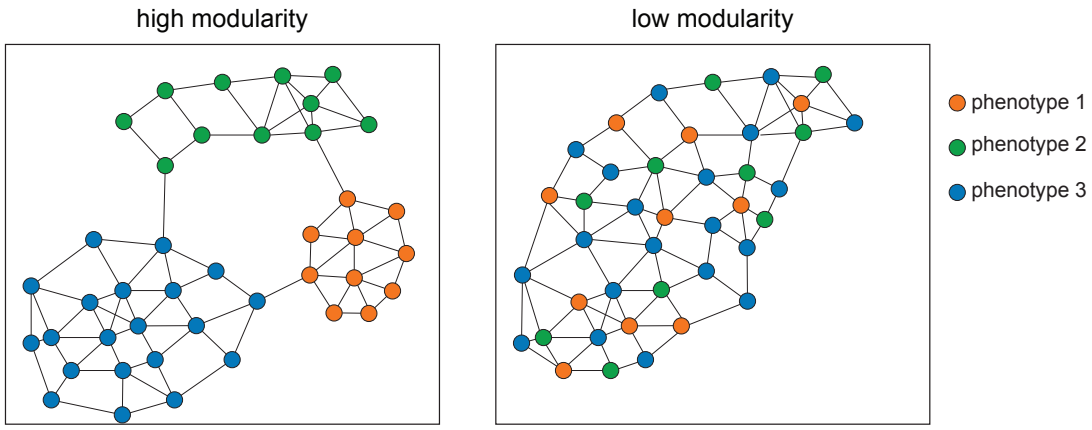
1.4.2 Graph-theoretic scores

An appealing alternative towards quantifying tumor heterogeneity is offered by exploiting graph-theoretic concepts that quantify different topological properties of the graph representation. A common measure in



Supplementary Figure 4: **Spatial statistics scores**. Ripley's K and its variance-stabilizing transform $L(r)$ quantify the extend of clustering, uniformity or dispersion as a function of radius r, exemplified here for a purple, orange and blue phenotype, respectively.

this category is **modularity** that has long been explored in biological networks and associated with their robustness [10]. Modularity captures the structure of a graph by quantifying the degree at which it can be divided into communities of the same label. In the context of tumor heterogeneity, modularity can be thought of as the degree of self-organization of the cells with the same phenotype into spatially distinct communities: as seen in Supplementary Figure 5, a graph of high modularity represents a tumor where connections between the cells within the same community are more dense than connections between cells of different communities.



Supplementary Figure 5: **Graph modularity.** Two hypothetical tumors and their corresponding cell graphs with different levels of graph modularity are shown. Although both tumors have the same phenotypic composition, the graph on the left is much more modular that the graph on the right.

Modularity is formally defined by the following equation [11]:

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \gamma \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) \tag{4}$$

where m is the number of connections, A is the adjacency matrix of graph G, k_i is the number of edges connected to node i (also known as *node degree*), γ is the resolution parameter, and $\delta(c_i, c_j)$ is 1 if nodes i and j belong to the same community and 0 otherwise. In practice, to compute the graph modularity the following equivalent equation is used [10]:

$$Q = \sum_{c=1}^{n} \left[\frac{L_c}{m} - \gamma \left(\frac{k_c}{2m} \right)^2 \right] \tag{5}$$

where L_c and k_c are the number of intra-community edges and the sum of degrees of all nodes in community c, respectively. ATHENA exploits networkx, a Python package that supports creation and analysis of complex graphs, which implements equation 5 to compute the graph modularity, included in ATHENA's .metrics submodule.

1.4.3 Information-theoretic scores

The quantification of the diversity in an ecosystem is a longstanding problem in ecology and a vast body of scientific literature has addressed the problem (for a thorough review of heterogeneity metrics applied to cancer research, see [3]). The application of the concepts developed in ecology to cancer research is straightforward, as there is a direct analogy between biological and tumor ecosystems. In general, the scores developed in ecology are functions of the number of species (*richness*) and their relative abundance, weighting each aspect differently depending on the score. The mathematical foundation of these scores is rooted in information theory, where similar concepts were employed to quantify information content or uncertainty of a communication system [12]. In the context of tumor heterogeneity, notable examples of information-theoretic scores include quantifying genetic heterogeneity using a variety of fluorescence in situ hybridization or immunofluorescence assays [13, 14, 15, 16, 17]. ATHENA currently implements the following information-theoretic scores, included in the .metrics submodule:

1. **Richness**: The most basic and intuitive heterogeneity score is richness S, which simply counts the number of observed cell subpopulations within a tumor sample, independently of their relative abundance. This score is equivalent to *tumor clonality*, commonly employed in genetic heterogeneity studies [3] to quantify the number of distinct clones in a tumor.

2. **Shannon index**: Shannon index H takes into consideration not only the number of cell subpopulations S present, but also their relative proportions. Assuming that $p(x_i)$ represents the probability of observing a cell subpopulation x_i, then it is defined as:

$$H = -\sum_{i=1}^{S} p(x_i) \log p(x_i) \tag{6}$$

   Another way to view Shannon entropy is as a quantification of the degree of 'surprise', in other words, how likely we are to guess the phenotype of a randomly observed cell from a tumor sample. This concept is visualized in Supplementary Figure 6, in which three different tumors with different phenotype distributions are shown: the more even the cell proportions, the more uncertain our prediction. Shannon entropy increases with richness and evenness, and reaches its maximal value when the cell subpopulation distribution is uniform.

3. **Simpson index**: Similarly, the Simpson index considers both richness and relative abundance and is defined as:

$$D = \sum_{i=1}^{S} p(x_i)^2 \tag{7}$$

   The Simpson index describes the probability of sampling the same phenotype twice from the tumor. In contrast to the Shannon index, the Simpson index decreases with increasing diversity. Furthermore, the Simpson index is sensitive to the abundance of the more dominant phenotype and can be regarded as a measure of dominance concentration [18].

4. **Hill Numbers**: The Hill numbers ${}^qD$ [18] constitute a family of heterogeneity scores under a unified mathematical expression:

$$ {}^q D = \left(\sum_{i=1}^{S} p(x_i)^q \right)^{1/(1-q)} \qquad 0 \leq q \leq \infty \tag{8} $$

It can be shown that richness S, Shannon index H and Simpson index D are all special cases of ${}^qD$:

$$ {}^0D = S \tag{9} $$

$$ {}^1D = \exp H \tag{10} $$

$$ {}^2D = \frac{1}{D} \tag{11} $$

As shown by [18], low values of q giving more emphasis to rare phenotypes. An interesting application of the Hill numbers in the biological domain can be found in immunology, where immune repertoires from different patients that are not directly comparable were mapped in a common diversity index space to predict immune status [19].

5. **Renyi entropy**: Similarly to the Hill numbers, Renyi entropy is a generalization of multiple entropic measures [20]. In the context of tumor heterogeneity, Renyi entropy has been previously used to quantify the diversity of the tumor lung TCR repertoires [21]. Renyi entropy is defined as:

$$ H_\alpha = \frac{1}{1-\alpha} \log \left(\sum_{i=1}^{S} p(x_i)^\alpha \right) \tag{12} $$

Similarly to the Hill numbers, it can be shown that, for different values of α, Renyi entropy is equivalent to other entropic expressions:

$$ H_0 = \log R \ \text{ (Hartley entropy)} \tag{13} $$

$$ H_1 = S \ \text{ (Shannon entropy)} \tag{14} $$

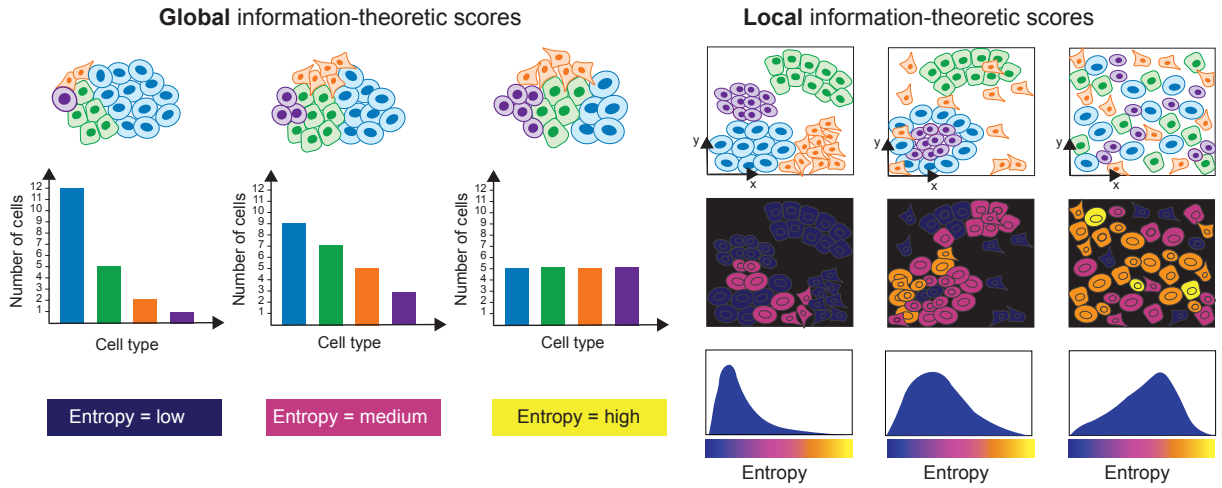
$$ H_2 = -\log \sum_{x \in X} p(x)^2 = -\log D \ \text{ (Collision entropy)} \tag{15} $$

$$ H_\infty = -\log \max_x p(x) \ \text{ (Min-entropy)} \tag{16} $$

6. **Rao's quadratic entropy**: The indices and quantification methods discussed so far consider both richness and relative abundance of phenotypes, but ignore the similarity of these phenotypes, i.e., how close they are in phenotypic space. Rao's quadratic entropy [22] accounts for that by incorporating a distance metric $d(x_i, x_j)$ and is defined as:

$$ Q = \sum_{i=1}^{S-1} \sum_{j=i+1}^{S} d(x_i, x_j) p(x_i) p(x_j) \tag{17} $$

In their original definition, all entropic scores described above do not take the spatial component into account when calculated on a whole tumor level. For this reason, ATHENA implements two flavors of these scores: a *global* flavor, in which the metrics are computed at a whole sample level using only the phenotype distribution, and a *local* flavor, in which the scores are computed at a single-cell level, using also the graph structure (see Supplementary Table 1 and Supplementary Figure 6). Specifically, when computing local scores, ATHENA iterates over all cells, and for each cell, computes the local entropy within its neighborhood. In this way, highly diverse regions where cells from multiple different phenotypes coexist can be highlighted, and, instead of computing a single entropy value as for the *global* flavor, a distribution of entropic values is returned.



Supplementary Figure 6: **Information-theoretic scores**. Global information-theoretic scores (e.g., Shannon, Simpson or quadratic entropy) take into consideration the number and relative proportions of phenotypes present in a tumor to compute tumor heterogeneity at a whole-tumor level, as shown here for three hypothetical tumors with different compositions. In contrast, local information-theoretic scores additionally account for the spatial positioning of these phenotypes, highlighting regions of high diversity, as shown here for three tumors that have the same phenotypic composition and thus same global entropy.

1.4.4 Cell interaction scores

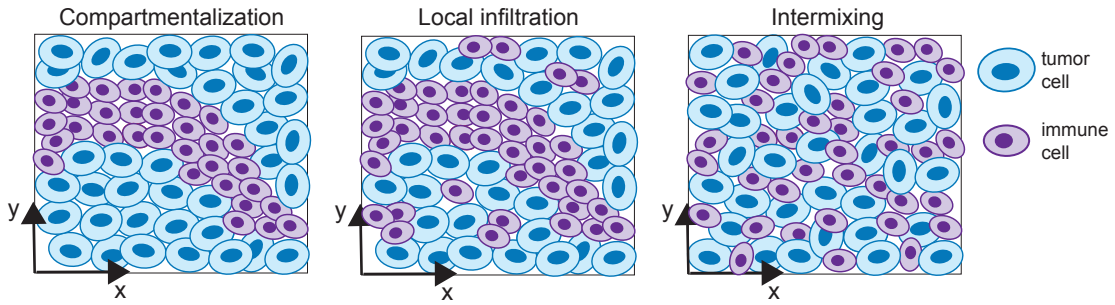
More sophisticated heterogeneity scores additionally consider cell-cell interactions by exploiting the cell-cell graph, where nodes encode cells, edges encode interactions, and each node is associated with a label that encodes the cell's phenotype. The cell interaction scores implemented in ATHENA's neigh submodule include:

1. **Infiltration score**: The infiltration score quantifies the degree to which a certain cell phenotype has penetrated among cells of another type. In cancer research, the degree to which immune cells infiltrate the tumor tissue is of particular interest and one can distinguish between different spatial organisations of this infiltration. For example, [23] distinguished the following tumor architectures: (i) *cold* tumors that feature a small number of immune cells, (ii) *mixed* tumors that have immune cells dispersed in the tumor mass and (iii) *compartmentalised* tumors, where tumor and immune cells self-cluster in separate, non-intermixed compartments. In [23], the tumor-immune mixing score is proposed, defined as the ratio of the number of immune-tumor interactions to the number of immune-immune interactions. The infiltration score implemented in ATHENA is a generalised form of the above ratio, formally defined as:

$$\frac{\text{number of cell type } i \text{ - cell type } j \text{ interactions}}{\text{number of cell type } i \text{ - cell type } i \text{ interactions}} \tag{18}$$

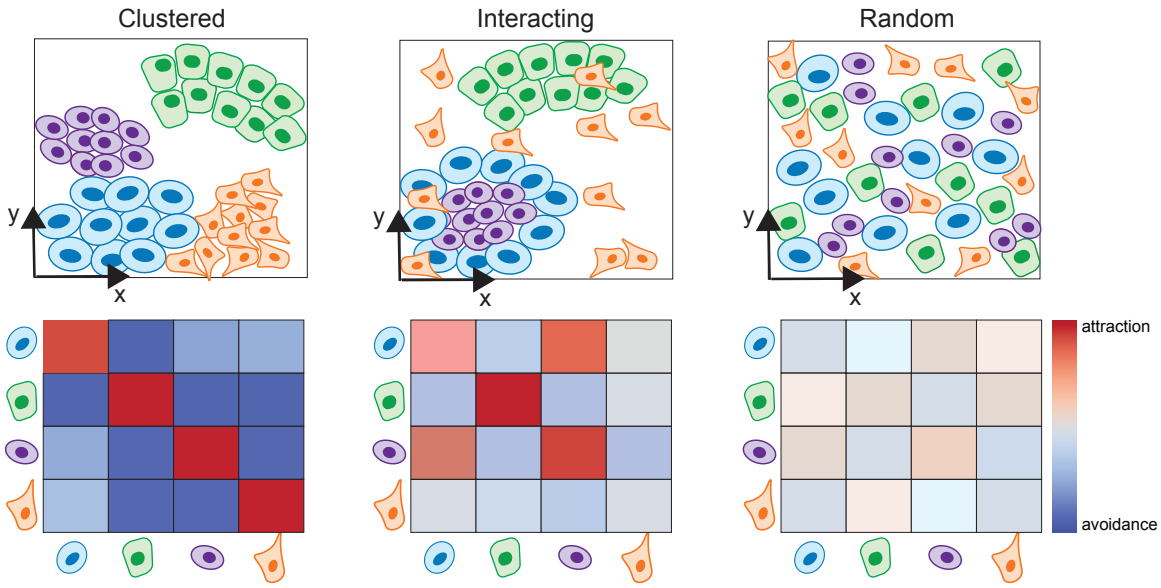
   where i, j the labels of any two cell phenotypes. As a consequence, the implementation of the score does not explicitly limit its application to immune-to-tumor infiltration, but is very flexible and allows the user to define any pairwise interaction, e.g., a specific immune subtype to the whole tumor, or even non-immune types of infiltration, should this be of interest. As obvious from the formula, the infiltration score is not defined in cases of samples that do not feature any cell type i - cell type i interactions.

   ATHENA implements two flavors of infiltration, a global one that returns an estimate at the whole-sample level, and a local one, where for each cell i the infiltration is computed on the sub-graph only containing all immediate neighbors of i.



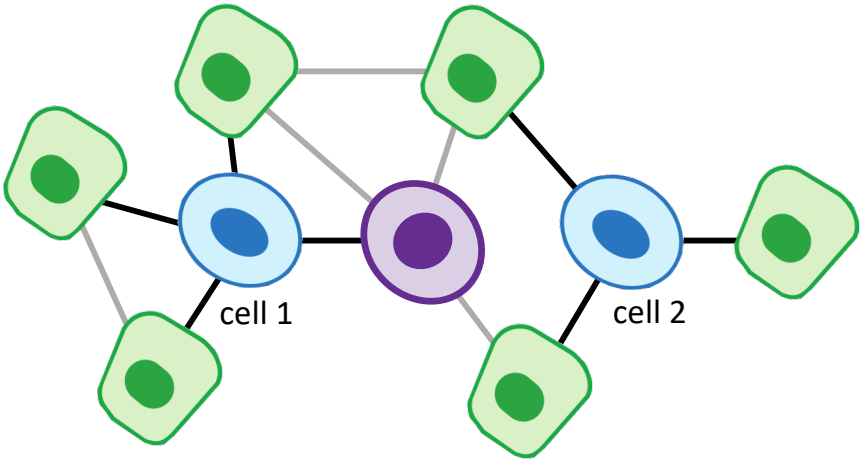
Supplementary Figure 7: **Infiltration score.** Three different tumors with three different infiltration patterns (left: compartmentalized, middle: local infiltration, right: intermixed).

2. **Neighborhood analysis score**: Neighborhood analysis [24] (also known as spatial enrichment analysis [23]) quantifies pairwise interactions between phenotypes by assessing whether they occur more or less often (attraction or avoidance, respectively) than expected by chance (see Supplementary Figure 8).



Supplementary Figure 8: **Neighborhood analysis.** Three different tumors of the same phenotypic composition that exhibit markedly different patterns of phenotype-phenotype interactions, as captured by the neighborhood analysis scores.

This is achieved by an iterative process, where, for each cell in the graph, neighboring cells are recorded, all pairwise interactions between different phenotypes are counted, and a global average interaction score across all cells of that phenotype is computed. Then, the significance of the observed pairwise interactions is assessed by a permutation test: cell labels in the graph are permuted, i.e. randomized while keeping their relative proportions fixed, and the pairwise interactions scores are recomputed for these random node phenotype assignments. The obtained H_0 distribution of interaction



	Counts cell 1	Counts cell 2	Classic flavor	HistoCAT flavor	Proportional flavor
Blue→Blue	0	0	0	0	0
Blue→Violet	1	0	0.5	1	0.125
Blue→Green	3	3	3	3	0.875

Supplementary Figure 9: **Different flavors of neighborhood analysis scores**. Working principles of the different flavors of neighborhood analysis, exemplified at two blue cells, highlighting how the global interaction strength of blue cells with green and violet cell types is computed.

scores is then used to compute a p-value for the observed interaction strength. ATHENA implements three flavors of this score (classic, histoCAT and proportional) that differ on how phenotype interactions are counted and averaged across all cells. An R implementation of the classic and histoCAT flavors is also available here: `https://bodenmillergroup.github.io/imcRtools/reference/countInteractions.html`. These different flavors are exemplified in the toy example of Supplementary Figure 9, where cell 1 has three green and one violet cell neighbors, and cell 2 only has three green cell neighbors.

Classic: In the classic flavor of the neighborhood analysis score, the average number of interactions between cells of the same phenotype is computed. In our example, this translates to an average score of 3 for the Blue→Green interaction, and an average of 0.5 for the Blue→Violet interactions.

HistoCAT: In the histoCAT flavor, the global average of interaction score is only computed across cells that actually show this interaction. In the scenario of Supplementary Figure 9, when computing the average interaction score Blue→Violet, only cell 1 is considered, since cell 2 does not have any violet cell neighbors. This leads to an inflated interaction value of 1 when compared to the classic flavor.

Proportional: Finally, the proportional flavor uses interaction frequencies instead of counts, i.e., divides the counts of pairwise interactions by all interactions a given cell has. In our toy example, cell 1 and cell 2 interact with violet cells at a frequency of 0.25 and 0, respectively, leading to a global average score of 0.125. Similarly, cell 1 and cell 2 interact with green cells at a frequency of 0.75 and 1, respectively, leading to a global average score of 0.875. We propose this approach as an alternative that normalizes the scores with respect to varying cell density. In contrast to the classic or the histoCAT flavor, the proportional flavor is not influenced by the number of cells in a sample and the score is bounded in the range of [0,1]. Furthermore, in addition to a p-value, we propose to compute the difference between the observed proportion of interactions obs_{ij} between cell type $i \rightarrow j$ and the randomised proportion of interactions rand_{ij}. This difference can be asymmetrical ($i \rightarrow j \neq j \rightarrow i$), is bounded in the range of [-1,1], and might be more suited as input for certain machine learning models.

Supplementary Tables

Score	Flavor	Input	Hyperparameter	Previously used
Spatial statistics				
Ripley's K	global	G,P	r, graph choice	[7, 8]
Graph-theoretic				
Modularity	global	G,P	γ, graph choice	ATHENA
Information-theoretic				
Richness	global	P	–	[25]
Richness	local	G,P	graph choice	ATHENA
Shannon index	global	P	–	[13, 15, 16]
Shannon index	local	G,P	graph choice	ATHENA
Simpson index	global	P	–	[13, 14, 16]
Simpson index	local	G,P	graph choice	ATHENA
Renyi entropy	global	P	α	[21]
Renyi entropy	local	G,P	α, graph choice	ATHENA
Hill numbers	global	P	q	[19]
Hill numbers	local	G,P	q, graph choice	ATHENA
Quadratic entropy	global	P	$D(x, y)$	[17, 26]
Quadratic entropy	local	G,P	$D(x, y)$, graph choice	ATHENA
Cell interaction				
Infiltration	global	G,P	graph choice	[23]
Infiltration	local	G,P	graph choice	ATHENA
Classic	global	G,P	graph choice	[24]
HistoCAT	global	G,P	graph choice	[24]
Proportion	global	G,P	graph choice	ATHENA

Supplementary Table 1: **Summary of all heterogeneity quantification scores included in ATHENA.** The flavor column indicates if the score is computed on a global (sample) level or on a local (observation) level. The input column specifies the input information used by the scores. Global scores use the phenotype distribution P and do not rely on spatial information. In contrast, scores that require a graph input use the spatial information encoded in the graph representation G. Results of some methods depend on additional hyperparameter choices (r: radius, γ: resolution parameter, α and q: parameteres used for Hill numbers and Renyi entropy, respectively, $D(x, y)$: a distance measure between single-cell profiles x and y.

References

[1] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, February 2018.

[2] Aric Hagberg, Pieter Swart, and Daniel S Chult. Exploring network structure, dynamics, and function using NetworkX. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[3] Aditya Kashyap, Maria Anna Rapsomaniki, Vesna Barros, Anna Fomitcheva-Khartchenko, Adriano Luca Martinelli, Antonio Foncubierta Rodriguez, Maria Gabrani, Michal Rosen-Zvi, and Govind Kaigala. Quantification of tumor heterogeneity: from data acquisition to metric generation. *Trends in Biotechnology*, 0(0), December 2021. Publisher: Elsevier.

[4] Erik Meijering and Gert van Cappellen. Biological image analysis primer. *Erasmus MC, Rotterdam*, 2006. Publisher: Citeseer.

[5] Pushpak Pati, Guillaume Jaume, Antonio Foncubierta-Rodríguez, Florinda Feroce, Anna Maria Anniciello, Giosue Scognamiglio, Nadia Brancati, Maryse Fiche, Estelle Dubruc, Daniel Riccio, Maurizio Di Bonito, Giuseppe De Pietro, Gerardo Botti, Jean-Philippe Thiran, Maria Frucci, Orcun Goksel, and Maria Gabrani. Hierarchical graph representations in digital pathology. *Medical Image Analysis*, 75:102264, January 2022.

[6] Philip M. Dixon. Ripley's K Function. In *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, Ltd, 2014. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118445112.stat07751.

[7] A. Francesca Setiadi, Nelson C. Ray, Holbrook E. Kohrt, Adam Kapelner, Valeria Carcamo-Cavazos, Edina B. Levic, Sina Yadegarynia, Chris M. van der Loos, Erich J. Schwartz, Susan Holmes, and Peter P. Lee. Quantitative, Architectural Analysis of Immune Cell Subsets in Tumor-Draining Lymph Nodes from Breast Cancer Patients and Healthy Lymph Nodes. *PLOS ONE*, 5(8):e12420, August 2010. Publisher: Public Library of Science.

[8] Chang Gong, Robert A. Anders, Qingfeng Zhu, Janis M. Taube, Benjamin Green, Wenting Cheng, Imke H. Bartelink, Paolo Vicini, Bing Wang, and Aleksander S. Popel. Quantitative Characterization of CD8+ T Cell Clustering and Spatial Heterogeneity in Solid Tumors. *Frontiers in Oncology*, 8, 2019. Publisher: Frontiers.

[9] Brian D. Ripley. The second-order analysis of stationary point processes. *Journal of applied probability*, 13(2):255–266, 1976. Publisher: Cambridge University Press.

[10] Sergio Antonio Alcalá-Corona, Santiago Sandoval-Motta, Jesús Espinal-Enríquez, and Enrique Hernández-Lemus. Modularity in Biological Networks. *Frontiers in Genetics*, 12:1708, 2021.

[11] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004. Publisher: APS.

[12] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948. Conference Name: The Bell System Technical Journal.

[13] Michalina Janiszewska, Lin Liu, Vanessa Almendro, Yanan Kuang, Cloud Paweletz, Rita A. Sakr, Britta Weigelt, Ariella B. Hanker, Sarat Chandarlapaty, Tari A. King, Jorge S. Reis-Filho, Carlos L. Arteaga, So Yeon Park, Franziska Michor, and Kornelia Polyak. In situ single-cell analysis identifies heterogeneity for PIK3CA mutation and HER2 amplification in HER2-positive breast cancer. *Nature Genetics*, 47(10):1212–1219, October 2015.

[14] So Yeon Park, Mithat Gönen, Hee Jung Kim, Franziska Michor, and Kornelia Polyak. Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype. *The Journal of Clinical Investigation*, 120(2):636–644, February 2010.

[15] Michael J. Gerdes, Yesim Gökmen-Polar, Yunxia Sui, Alberto Santamaria Pang, Nicole LaPlante, Adrian L. Harris, Puay-Hoon Tan, Fiona Ginty, and Sunil S. Badve. Single-cell heterogeneity in ductal carcinoma in situ of breast. *Modern Pathology*, 31(3):406–417, March 2018. Number: 3 Publisher: Nature Publishing Group.

[16] Yul Ri Chung, Hyun Jeong Kim, Young A. Kim, Mee Soo Chang, Ki-Tae Hwang, and So Yeon Park. Diversity index as a novel prognostic factor in breast cancer. *Oncotarget*, 8(57):97114–97126, September 2017.

[17] Steven J. Potts, Joseph S. Krueger, Nicholas D. Landis, David A. Eberhard, G. David Young, Steven C. Schmechel, and Holger Lange. Evaluating tumor heterogeneity in immunohistochemistry-stained breast cancer tissue. *Laboratory Investigation*, 92(9):1342–1357, September 2012. Number: 9 Publisher: Nature Publishing Group.

[18] Mark O. Hill. Diversity and evenness: a unifying notation and its consequences. *Ecology*, 54(2):427–432, 1973. Publisher: Wiley Online Library.

[19] Victor Greiff, Pooja Bhat, Skylar C. Cook, Ulrike Menzel, Wenjing Kang, and Sai T. Reddy. A bioinformatic framework for immune repertoire diversity profiling enables detection of immunological status. *Genome Medicine*, 7(1):49, May 2015.

[20] Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 547–561. University of California Press, 1961.

[21] Kroopa Joshi, Marc Robert de Massy, Mazlina Ismail, James L. Reading, Imran Uddin, Annemarie Woolston, Emine Hatipoglu, Theres Oakes, Rachel Rosenthal, Thomas Peacock, Tahel Ronel, Mahdad Noursadeghi, Virginia Turati, Andrew J. S. Furness, Andrew Georgiou, Yien Ning Sophia Wong, Assma Ben Aissa, Mariana Werner Sunderland, Mariam Jamal-Hanjani, Selvaraju Veeriah, Nicolai J. Birkbak, Gareth A. Wilson, Crispin T. Hiley, Ehsan Ghorani, José Afonso Guerra-Assunção, Javier Herrero, Tariq Enver, Sine R. Hadrup, Allan Hackshaw, Karl S. Peggs, Nicholas McGranahan, Charles Swanton, Sergio A. Quezada, and Benny Chain. Spatial heterogeneity of the T cell receptor repertoire reflects the mutational landscape in lung cancer. *Nature Medicine*, 25(10):1549–1559, October 2019. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 10 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Biological sciences;Immunoediting Subject_term_id: biological-sciences;immunoediting.

[22] C. Radhakrishna Rao. Diversity and dissimilarity coefficients: a unified approach. *Theoretical population biology*, 21(1):24–43, 1982. Publisher: Elsevier.

[23] Leeat Keren, Marc Bosse, Diana Marquez, Roshan Angoshtari, Samir Jain, Sushama Varma, Soo-Ryum Yang, Allison Kurian, David Van Valen, Robert West, Sean C. Bendall, and Michael Angelo. A Structured Tumor-Immune Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam Imaging. *Cell*, 174(6):1373–1387.e19, September 2018.

[24] Denis Schapiro, Hartland W. Jackson, Swetha Raghuraman, Jana R. Fischer, Vito RT Zanotelli, Daniel Schulz, Charlotte Giesen, Raúl Catena, Zsuzsanna Varga, and Bernd Bodenmiller. histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nature methods*, 14(9):873, 2017. Publisher: Nature Publishing Group.

[25] Johanna Wagner, Maria Anna Rapsomaniki, Stéphane Chevrier, Tobias Anzeneder, Claus Langwieder, August Dykgers, Martin Rees, Annette Ramaswamy, Simone Muenst, Savas Deniz Soysal, Andrea Jacobs, Jonas Windhager, Karina Silina, Maries van den Broek, Konstantin Johannes Dedes, Maria Rodríguez Martínez, Walter Paul Weber, and Bernd Bodenmiller. A Single-Cell Atlas of the Tumor and Immune Ecosystem of Human Breast Cancer. *Cell*, 177(5):1330–1345.e18, May 2019.

[26] Daniel M. Spagnolo, Yousef Al-Kofahi, Peihong Zhu, Timothy R. Lezon, Albert Gough, Andrew M. Stern, Adrian V. Lee, Fiona Ginty, Brion Sarachan, D. Lansing Taylor, and S. Chakra Chennubhotla. Platform for Quantitative Evaluation of Spatial Intratumoral Heterogeneity in Multiplexed Fluorescence Images.

Cancer Research, 77(21):e71–e74, November 2017. Publisher: American Association for Cancer Research Section: Focus on Computer Resources.