Spatial analysis for highly multiplexed imaging data to identify tissue microenvironments

Authors

Ellis Patrick^{1,2,*}, Nicolas P. Canete^{2,3}, Sourish S. Iyengar¹, Andrew N. Harman^{2,5}, Greg T. Sutherland^{4,5} and Pengyi Yang^{1,4,6}

Affiliations

¹ School of Mathematics and Statistics, The University of Sydney, Sydney, New South Wales, Australia

² Centre for Cancer Research, Westmead Institute for Medical Research, The University of Sydney, Sydney, New South Wales, Australia

³ Sydney Medical School, Faculty of Medicine and Health, The University of Sydney, Sydney, New South Wales, Australia

⁴ Charles Perkins Centre, The University of Sydney, Sydney, New South Wales, Australia

⁵ School of Medical Sciences, Faculty of Medicine and Health, The University of Sydney, Sydney, New South Wales, Australia

⁶ Computational Systems Biology Group, Children's Medical Research Institute, Faculty of Medicine and Health, The University of Sydney, Westmead, NSW, Australia

*Corresponding author. Email: ellis.patrick@sydney.edu.au

Abstract

Highly multiplexed *in situ* imaging cytometry assays have made it possible to study the spatial organisation of numerous cell types simultaneously. We have addressed the challenge of quantifying complex multi-cellular relationships by proposing a statistical method which clusters local indicators of spatial association. Our approach successfully identifies distinct tissue architectures in datasets generated from three state-of-the-art high-parameter assays demonstrating its value in summarising the information-rich data generated from these technologies.

Main

Understanding the interplay between different cells, cell sub-types and their immediate environment is critical for understanding the mechanisms by which cells organise themselves within tissues and their dysfunctions in the context of human diseases. Recent advances in high-parameter imaging cytometry technologies which include among others¹ High-definition Spatial Transcriptomics², seqFISH+³, CODEX⁴, t-CyCIF⁵, MIBI-TOF⁶ and IMC⁷ have fundamentally revolutionised our ability to observe these complex cellular relationships. Where standard immunohistochemistry protocols only allow the accurate visualisation of cells that can be uniquely characterised by two or three surface proteins, these cutting-edge high-parameter imaging cytometry technologies characterise cells with tens to tens of thousands of RNA or proteins enabling precise classification of cell sub-types and hence providing an unprecedented depiction of cellular heterogeneity in the spatial context in a tissue. The increase in size and high-dimensional nature of the images necessitates analytical tools that are more complex, computationally efficient, and require an increased need for interpretability than those currently used in standard imaging cytometry.

As low parameter (two or three) imaging cytometry protocols are well established, so too are the analytical techniques needed to analyse the data they produce. The two most widely used techniques for object-based analysis of cell localisation are nearest neighbour analysis and Poisson point process models⁸. Nearest neighbour analysis can be used to test if two cell types tend to be more closely located than would be expected by chance. In a k-nearest neighbour analysis, for each cell in cell type c_1 , the k closest cells are identified and then the number of another cell type c_2 in this set are counted. These counts for each cell-region are then averaged and compared to the averages of cell-label permuted experiments to quantify significance. Alternatively, point process models have been used for decades in fields such as ecology, geology and astronomy to identify the localisation of objects. The K-function can be used to quantify and visualise a point process model. Over a range of radii, r , the K-function is a monotonic quantification of the average number of c_2 cells within r distance of each c_1 cell and adjusts for the expected number of cells in each cell region. Under certain assumptions statistical significance can be assessed for a Poisson point process model without the use of permutations.

There is a substantial need to develop analytical approaches for simultaneously identifying relationships between multiple cell types. Given the novelty of the technologies, analytical approaches for tackling this issue are in active development. Methods of note include Spatial Variance Component Analysis (SVCA) which addresses the problem by analysing images at the marker level, but amongst other influential effects, can quantify the changes in marker abundance explained by cell-cell interactions⁹. Alternatively, Goltsev *et al.* proposed an approach identifying 100 cliques of cell types from a nearest neighbour analysis showing that these captured histologically

distinct regions of cell type composition⁴. Cellular neighbourhoods are identified by CytoMAP¹⁰ by clustering cell compositions of rasterised images defined by either scanning across the image or across cells and Giotto¹¹ identifies spatial domains using a hidden Markov random field.

Here we propose a novel methodology implemented in R for identifying consistent spatial organisation of multiple cell types in an unsupervised way. Our method leverages established, and well-defined metrics to quantify point-process models. In short, our method clusters local indicators of spatial association (LISA) curves to enable the characterisation of multiple interactions between cell types in contrast to traditional pairwise analysis (**Figure 1**). LISA curves are a localised summary of a K-function from a Poisson point process model and have been used in varied contexts to identify landmines¹² or clusters of forest fires and earthquakes¹³. In the following, we demonstrate that clustering these curves from multiple cell type cross-comparisons is a robust and stable method for providing a high-level summary of cell type localisation in high-parameter imaging cytometry data.

Our method, lisaClust, can identify distinct and well-defined tissue compartments in the spleen. The co-detection by indexing (CODEX) high-parameter imaging protocol was recently used to profile 26 immune cell types in murine spleens⁴. Applying the elbow method to estimate the number of clusters would suggest that the spatial relationships between the 26 cell types could be used to cluster the cells into three groups with distinct spatial ordering (**Supplementary Figure 1**). However, by clustering the cells into four groups we detected the four classic splenic compartments: red pulp, B cell follicles, PALS (periarteriolar lymphoid sheath), and marginal zone (**Figure 2a**). The putative B cell follicle was enriched for B-cells and follicular dendritic cells, the putative PALS was enriched for CD4+ and CD8+ T cells, the putative marginal zone enriched for marginal zone macrophages and the putative red pulp region enriched for plasma cells (**Figure 2b**). The unsupervised identification of these tissue compartments was robust to the choice of clustering method with k-means clustering, hierarchical clustering, Clustering Large Applications (CLARA)¹⁴ and Self-Organising Maps (SOM)¹⁵ all identifying similar regions (**Supplementary Figure 2**).

The density of cells across a tissue can influence the types of regions identified by lisaClust. Depending on the scenario, identifying spatial domains which contain complex cellular interactions might be preferred to identifying regions which are representative of dense tissue architecture. This contrast can be observed in a High-definition Spatial Transcriptomics (HDST) dataset which profiled the spatial organisation of 63 cell types in the murine olfactory bulb (**Figure 2c**). When lisaClust was used to identify three distinct regions (**Figure 2d**), these regions primarily reflected the density of cells in the tissue (**Figure 2e**) with one region capturing the densest portion that is the inner granule cell layer (**Figure 2d, Supplementary Figure 3a**). However, statistical models which account for strong spatial structure are well defined. By modelling the spatial organisation of cells as an inhomogeneous Poisson point process and calculating the corresponding LISA curves, lisaClust was instead able to demarcate the whole granule cell layer and the subependymal zone (**Figure 2f, Supplementary Figure 3b**). This illustrates the importance of the operator being cognisant about whether identified regions will be influenced by cell density and that lisaClust is flexible to this decision.

Finally, lisaClust can identify spatial ordering of cell types that are associated with a measured outcome across samples. Imaging Mass Cytometry (IMC) has been used to profile 16 cell types in pancreatic islets from individuals with recent onset type 1 diabetes, long-term type 1 diabetes and non-diabetics⁴. After applying lisaClust to identify four distinct regions, the overall area of these

regions was used to quantify the amount or level of specific complex cell type interactions in each sample. Two of the regions were representative of islets and were highly enriched for alpha cells, delta cells and gamma cells while one region also contained beta cells and another no beta cells at all (**Figure 2g**). Beta cells produce insulin and so, unsurprisingly, the regions that contain the beta cells were present in non-diabetics but not in long-term type 1 diabetics (**Figure 2h**). While simpler analysis strategies might have identified that the proportion of beta cells was different between diabetics and non-diabetics, they would not have placed this difference into the broader context of their loss from the islet. This approach demonstrates an alternative to multiple pairwise localisation tests and is a viable step towards a multivariate analysis procedure.

Overall, clustering local indicators of spatial association is an effective approach for quantifying high-dimensional spatial ordering in highly multiplexed *in situ* imaging cytometry data. However, it does have limitations as three parameters need to be chosen by the operator. In general, like the elbow method which we used, many methods exist for optimising the number of clusters to choose when clustering. We have not evaluated these here given a lack of ground truth and that, in our opinion, the number of clusters chosen should be contextual. The range of radii over which to calculate the curves also needs to be decided by the operator but clustering curves, instead of a single distance, appears to lessen the impact of this choice¹⁶ (Supplementary Figure 4). Further, the choice of whether to use an inhomogeneous Poisson point process model, or not, can alter the types of regions identified (**Figure 2e**, **Figure 2f**) and should be selected purposefully.

We have illustrated that lisaClust can be applied to identify tissue compartments and prioritise distinct cellular microenvironments in multiple imaging assays that profile cellular composition at a single-cell resolution. The method is available as an open-source R package (<https://bioconductor.org/packages/release/bioc/html/lisaClust.html>) and as an online shiny application <https://shiny.maths.usyd.edu.au/lisaClust/>.

Acknowledgements

We thank the scientific community for proactively making their data and code publicly available. This work has been partly supported by The University of Sydney and an Australian Research Council Discovery Early Career Researcher Award (DE200100944) funded by the Australian Government.

Contributions

All authors contributed to the interpretation of results and writing of the manuscript. E.P and N.P.C conceived the project with E.P and S.S.I performing the main analysis and E.P and P.Y generating the figures.

Methods

lisaClust

We propose Local Indicators of Spatial Association Clustering (lisaClust) as an algorithm for identifying distinct tissue microenvironments. The identified tissue microenvironments are regions of tissue that are enriched for combinations of certain cell types. Practically, lisaClust is a dimension reduction strategy which facilitates identifying and summarising complex spatial organisation of multiple cell types.

Local indicators of spatial association.

lisaClust is an object-based spatial analysis technique where each cell is represented as a labelled point in a two-dimensional plane. As such, it is assumed that any input data has undergone single-cell segmentation and cell type classification. The core of the approach is clustering localised summaries of multitype Poisson point process models which quantify the spatial localisation of pairs of cell types.

By modelling cells from each image as a multi-type Poisson point process, the spatial localisation between two cell types can be quantified with a K-function¹⁷,

$$\hat{K}_{ij}(r) = \frac{|W|}{n_i n_j} \sum_{x_l \in X^{(i)}} \sum_{x_k \in X^{(j)}} \mathbf{1}\{|x_l - x_k| \leq r\} e_{ij}(r),$$

where $\hat{K}_{ij}(r)$ summarises the degree of localisation of cell type j with cell type i within radius r , n_i and n_j are the number of cells of type i and j , x_m are the x-y coordinates of cell m , $|W|$ is the image area and $e_{ij}(r)$ is an edge correcting factor. The K-function can be interpreted as the average number of cells of type j within a distance r away from each cell of type i .

The L-function is a variance-stabilised K-function given by

$$\hat{L}_{ij}(r) = \sqrt{\frac{\hat{K}_{ij}(r)}{r}}.$$

The expected value of $L_{ij}(r)$ is r and so the centered L-function $\hat{L}_{ij}(r) - r$ will have constant variance and a mean of zero.

To calculate local indicators of spatial association, the contribution of each individual cell to the multi-type K-functions of cell type j is quantified as follows;

$$\hat{K}_j(r, x_m) = \frac{|W|}{n_j} \sum_{x_k \in X^{(j)}} \mathbf{1}\{|x_m - x_k| \leq r\} e_{ij}(r).$$

and then clustered on $\hat{L}_j(r, x_m) - r$. This measure is greater than zero if there are more cells of type j near cell m than expected and less than zero if there are less than expected.

Weights for tissue inhomogeneity.

Tissue typically has non-uniform structure which means that assuming the point process has a homogeneous intensity over the range of the image is inappropriate. To better summarise the degree of localisation in images with strong structure, the local contributions of cells to the inhomogeneous K-function is also calculated

$$\hat{K}_{ij}^{inhom} = \frac{1}{|W|} \sum_{x_l \in X^{(i)}} \sum_{x_k \in X^{(j)}} \frac{\mathbf{1}\{|x_l - x_k| \leq r\} e(x_l, x_k)}{\lambda(x_l) \lambda(x_k)},$$

where $\lambda(x_m)$ is the expected density of cells at cell x_m . $\lambda(x_m)$ is estimated using a Gaussian kernel with standard deviation 20.

Clustering

For the key results in this manuscript, k-means clustering is used to group the cells by their local indicators of spatial association using the *kmeans* function in the *stats* package in R. A fast implementation of hierarchical clustering from the *fastcluster* package¹⁸ is used for hierarchical clustering, Self-Organising Maps (SOM) are generated using the *kohonen* package and Clustering Large applications (CLARA) is implemented using the *cluster* package..

Datasets

lisaClust is applied to three highly multiplexed *in situ* imaging cytometry datasets. Each of these datasets was generated with a different technology.

Co-detection by indexing (CODEX)

CODEX (CO-Detection by indEXing) is a highly multiplexed cytometric imaging approach that uses antibodies conjugated to oligonucleotide sequences⁴. After staining a tissue sample with an antibody panel, dye labelled reporters with complementary sequences can be iteratively imaged and removed to quantify protein abundance on cells. Segmented data of a “BALBc-1” spleen sample taken from a normal BALB/c mouse was downloaded from <https://data.mendeley.com/datasets/zinpw8m5b/1>. A 30-antibody panel had been used to image 81693 cells and these were annotated into 26 cell types by the authors.

High-definition Spatial Transcriptomics (HDST)

HDST is an assay which captures RNA from tissue sections on a dense, spatially-barcoded bead array². The beads can measure the expression of thousands of RNA at 2 µm resolution making it possible to measure RNA expression at single cell resolution. The segmented data was downloaded from https://static-content.springer.com/esm/art%3A10.1038%2Fs41592-019-0548-y/MediaObjects/41592_2019_548_MOESM7_ESM.xlsx of a main olfactory bulb from a C57BL/6J mouse. This single image contained 69143 cells, and these were annotated into 63 cell types by the authors.

Imaging Mass Cytometry (IMC)

IMC is a highly multiplexed cytometric imaging approach that uses antibodies conjugated to metal isotopes⁷. After staining a tissue section with the metal-tagged antibodies, the tissue is ablated using a laser with a 1 µm spot size, the plumes of tissue matter are aerosolised, atomised, and ionised, and directed through a time-of-flight mass spectrometer for isotope quantification. The segmented data of 845 images of pancreatic islets from four non-diabetic human donors, four donors with recent onset type 1 diabetes and four donors with long duration type 1 diabetes were downloaded from <https://data.mendeley.com/datasets/cydmwsfztj/1>. A 35-antibody panel had been used to classify 1776974 cells into 16 distinct cell types by the authors.

Code availability

The code for generating the figures in this manuscript is available on <https://github.com/ellispatrick/lisaClustPaper>

References

1. Lewis, S. M. *et al.* Spatial omics and multiplexed imaging to explore cancer biology. *Nat Methods* 1–16 (2021).
2. Vickovic, S. *et al.* High-definition spatial transcriptomics for in situ tissue profiling. *Nat Methods* **16**, 987–990 (2019).
3. Eng, C.-H. L. *et al.* Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* **568**, 235–239 (2019).
4. Goltsev, Y. *et al.* Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging. *Cell* **174**, 968-981.e15 (2018).
5. Lin, J.-R. *et al.* Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes. *eLife* **7**, e31657 (2018).
6. Keren, L. *et al.* MIBI-TOF: A multiplexed imaging platform relates cellular phenotypes and tissue structure. *Science Advances* **5**, eaax5851 (2019).
7. Giesen, C. *et al.* Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature Methods* **11**, 417–422 (2014).
8. Baharlou, H., Canete, N. P., Cunningham, A. L., Harman, A. N. & Patrick, E. Mass Cytometry Imaging for the Study of Human Diseases—Applications and Data Analysis Strategies. *Front Immunol* **10**, (2019).
9. Arnol, D., Schapiro, D., Bodenmiller, B., Saez-Rodriguez, J. & Stegle, O. Modelling cell-cell interactions from spatial molecular data with spatial variance component analysis. *bioRxiv* 265256 (2018).
10. Stoltzfus, C. R. *et al.* CytoMAP: A Spatial Analysis Toolbox Reveals Features of Myeloid Cell Organization in Lymphoid Tissues. *Cell Reports* **31**, 107523 (2020).
11. Dries, R. *et al.* Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome Biology* **22**, 78 (2021).

12. Cressie, N. & Collins, L. B. Patterns in spatial point locations: Local indicators of spatial association in a minefield with clutter. *Naval Research Logistics (NRL)* **48**, 333–347 (2001).
13. Mateu, J., Lorenzo, G. & Porcu, E. Features detection in spatial point processes via multivariate techniques. *Environmetrics* **21**, 400–414 (2010).
14. Clustering Large Applications (Program CLARA). in *Finding Groups in Data* 126–163 (John Wiley & Sons, Ltd, 1990).
15. Kohonen, T. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **43**, 59–69 (1982).
16. Canete, N. P. *et al.* *spicyR: Spatial analysis of in situ cytometry data in R*. 2021.06.07.447307 <https://www.biorxiv.org/content/10.1101/2021.06.07.447307v1> (2021).
17. Ripley, B. D. The Second-Order Analysis of Stationary Point Processes. *Journal of Applied Probability* **13**, 255–266 (1976).
18. Müllner, D. fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python. *Journal of Statistical Software* **53**, 1–18 (2013).
19. Van Gassen, S. *et al.* FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry A* **87**, 636–645 (2015).

Figure legends

Figure 1 – Clustering local indicators of spatial association

The workflow that lisaClust uses to identify regions of tissue with similar localisation patterns of cells contains multiple key steps. Firstly, cells are treated as objects and assigned coordinates in an x-y space. Second, distances between all cells are calculated and then, by modelling the cells as a multi-type Poisson point process, the distances are used to calculate local indicators of spatial association (lisa). These lisa curves are local L-functions summarising the spatial association between each cell and a specific cell type over a range of radii, r . The lisa curves are calculated for each cell and celltype and then clustered to assign a region label for each cell.

Figure 2 – lisaClust is an effective approach for quantifying high-dimensional spatial ordering in high-parameter *in situ* imaging cytometry data.

a) In a sample of spleen tissue imaged with CODEX, lisaClust is used to cluster 81693 cells into four distinct regions highlighted by different hashing patterns. The four regions are representative of the red pulp, lymph follicles, PALS and marginal zone. Specific cell types that were enriched in the red pulp (red), lymph follicles (green), PALS (blue) and marginal zone (yellow) are plotted. **b)** a heatmap is used to quantify the enrichment of the 26 cell types in each region highlighted in a). **c)** HDST is used to profile 69143 cells with 63 cell types in a murine olfactory bulb (OB). Cells are plotted in x-y coordinates (gray) with *OBNBL1+* neurons (green) and *OBINH1+* neurons (orange) highlighted. **d)** density of the OB cells with high density in pink and low density in blue. **e)** Estimates using a homogenous local L-function capture OB regions with different cell density and **f)** regions estimated from an inhomogenous local L-function outline the OB granule cell layer and the subependymal zone. **g)** lisaClust is used to identify four regions contained within 845 IMC images of pancreatic islets and a heatmap summarises the enrichment of 16 cell types in each region. **h)** The proportion of cells from region r2, which contain alpha cells, delta cells and gamma cells but few beta cells, is higher in images from long duration type 1 diabetes individuals than non-diabetics.

Figure 1

a