SpatialQPFs: An R package for deciphering cell-cell spatial relationship

Xiao Li[1,*]

[1]Data Analytics and Imaging, Personalized Healthcare, Genentech, Inc., South San Francisco, CA, 94080, USA.
[*]To whom correspondence should be addressed: xiao.li.xl2@roche.com

ABSTRACT

Understanding the spatial dynamics within tissue microenvironments is crucial for deciphering cellular interactions and molecular signaling in living systems. These spatial characteristics govern cell distribution, extracellular matrix components, and signaling molecules, influencing local biochemical and biophysical conditions. Decoding these features offers insights into physiological processes, disease progression, and clinical outcomes. By elucidating spatial relationships between cell types, researchers uncover tissue architecture, cell communication networks, and microenvironment dynamics, aiding in the identification of biomarkers and therapeutic targets. Digital pathology imaging, including Hematoxylin and Eosin (H&E) staining, provides high-resolution histological information that offer intricate insights into cell-cell spatial relationships with greater details. However, current methods for capturing cell-cell spatial interactions are constrained by either methodological scopes or implementations restricted to script-level access. This limitation undermines generalizability and standardization, crucial for ensuring reproducibility. To address these limitations, we introduce *SpatialQPFs*, an extendable R package designed for extraction of interpretable spatial features from digital pathology images. By leveraging segmented cell information, our package provides researchers with a comprehensive toolkit for applying a range of spatial statistical methods within a stochastic process framework which includes analysis of point pattern data, areal data, and geostatistical data. This allows for a thorough analysis of cell spatial relationships, enhancing the depth and accuracy of spatial insights derived from the tissue, thereby empowering researchers to conduct comprehensive spatial analyses efficiently and reproducibly.

Introduction

Understanding the spatial characteristics within tissue microenvironments is crucial for unraveling the intricate dynamics of cellular interactions and molecular signaling within living systems. These spatial patterns dictate the distribution of cells, extracellular matrix components, and signaling molecules, shaping the local biochemical and biophysical environment. Deciphering these spatial characteristics provides critical insights into various physiological and pathological processes, including tissue development, homeostasis, disease progression and ultimately associates with patient clinical outcomes[1–4]. By discerning spatial relationships between cell types, researchers are offered valuable insights into tissue architecture, cell-cell communication networks, and microenvironment dynamics, facilitating the identification of spatially defined biomarkers and therapeutic targets for precision medicine approaches[5–7].

With the advance of imaging technology, digital pathology imaging of tumor tissue slides, such as those stained with Hematoxylin and Eosin (H&E) is becoming a routine clinical procedure for cancer management. This process produces massive information that capture histological details in high resolution. Thanks to the recent advances in artificial intelligence, automatic detection of different cell identities and corresponding spatial coordinates are made possible at large scale. Applying principles and advanced quantitative methods on these captured cell data, e.g. spatial statistics, can suggest novel solutions to fulfill this need. Quantitative analysis on these cellular-level data has produced encouraging findings[5,8–13]. However, all these previous efforts have been limited by their singular focus on a particular method for executing spatial analysis. Additionally, the code implementations are confined to script-level access tailored for specific data use cases, lacking the generalizability and standardization provided by a library, which is essential for reproducibility.

To overcome these limitations, we present *SpatialQPFs* (**S**patial **Q**uantitative **P**athology **F**eature**s**), a comprehensive and extendable R[14] package designed for deciphering spatial characteristics within/between cell types in the tissue microenvironment. By leveraging advanced imaging analysis techniques like tumor region-of-interest (ROI) identification, cell segmentation and classification, it becomes feasible to convert the raw, information-rich digital pathology images into a structured tabular format containing cell-level details, such as cell coordinates and identities. *SpatialQPFs* then take this cell level data as input and organizes it as point process data, areal data and geostatistical data, respectively, where a collection of spatial statistics methodology can be employed to extract human interpretable quantitative pathology features, empowering researchers to conduct comprehensive spatial analyses efficiently and reproducibly.

SpatialQPFs R Package

The *SpatialQPFs* package is written in R[14], an open-source, functional language and environment designed for statistical exploration and data visualization. Users can develop and customize the provided functions in the *SpatialQPFs* source code for various purposes and use cases.

The package primarily focuses on examining cell-cell spatial relationships. However, its functions are also relevant in diverse clinical contexts, like analyzing spatial patterns of brain lesions[15,16]. We intend to introduce further specialized functions for such applications in forthcoming updates.

The major functions of the *SpatialQPFs* package, detailed in Table 1, are geared towards streamlining the analysis of structured tabular data at the single-cell level. This is achieved through the organization of data into various spatial data types, each accompanied by a variety of visualization options.

Table 1. Primary SapatialQPFs functions (see package manual for further details).

Function	Description
Data_Vis()	Visualizes the input cell level data, including raw spatial map of individual cells and smoothed spatial density map
Point_pattern_data_uni()	Main function to generate spatial features using point process data (single cell type).
Point_pattern_data_bi()	Main function to generate spatial features using point process data (pairwise cell types).
Point_pattern_data_ITLR()	This is a function that discriminates a target cell population into 2 subgroups, in spatial relationships with the reference cell population. In case of lymphocytes and tumor cells, the lymphocytes are discriminated into intra-tumoral lymphocytes and adjacent-tumoral lymphocytes.
Areal_data()	Main function to generate spatial features using areal data.
Geostatistics_data()	Main function to generate spatial features using geostatistical data.

Data preparation

Feature extraction from digital images begins by computing maps for individual cell types, utilizing the center of each cell to create a cell mask. This process involves several image preprocessing steps, including but not limited to quality control measures such as artifact removal, identifying tumor regions of interest (ROIs), and performing cell classification and segmentation. These steps ensure that the data is clean and accurately represents the biological structures of interest.

Once the preprocessing is complete, the resulting cell mask from cell classification and segmentation step is converted to a structured tabular format, which includes the coordinates in the x- and y-planes and the identified cell types. This tabular data provides a comprehensive and organized representation of the spatial distribution and characteristics of the cells within the image.

These crucial steps precede the extraction of spatial features by *SpatialQPFs*, which analyze the spatial relationships and patterns among the cells. The tabular cell-level data and its derived spatial density map can be visualized by calling the `Data_Vis()` function. This visualization function allows researchers to explore and interpret the spatial organization of the cells, facilitating a deeper understanding of the spatial layout and interactions within the tissue microenvironment.

Point process data

Spatial point pattern analysis is a vital technique in spatial statistics, used to examine the distribution of point objects such as cell nuclei and their spatial interplay. This method is crucial for understanding how cells are organized within a given tissue, which can provide insights into various biological processes and disease mechanisms.

To leverage point process methodologies in characterizing the spatial distribution of a specific cell type, *SpatialQPFs* provides the `Point_pattern_data_uni()` function. This function employs advanced techniques such as Ripley's K-function, G-function, and Clark and Evans nearest neighbor analysis to determine whether the distribution of the point process object is random, clustered, or uniform[17,18]. These analyses help in identifying patterns within the spatial arrangement of cells, which can be critical for interpreting biological phenomena. Furthermore, to quantify the spatial interplay between two cell identities, such as lymphocytes and tumor cells, *SpatialQPFs* includes the `Point_pattern_data_bi()` and `Point_pattern_data_ITLR()` functions. These functions utilize a variety of sophisticated methods, including the cross-type Ripley's K-function, differences in Ripley's K-function, cross-type G-function, Pair correlation function, Mark correlation function, Mark connection function, Marcon and Puech's M function, and the cell population discrimination analysis. These methods provide detailed insights into the spatial relationships between different cell populations within the tissue, revealing interactions that may be crucial for understanding tumor biology and immune response[10,11,17,19–21]. The features derived from these point process methods are detailed in Supplementary material Section 1.

Areal data

To apply spatial statistics methods for areal data effectively, a conversion process is required, involving the partitioning of the study area into a grid of subregions or lattices, each annotated with its corresponding geographic location, typically denoted by the coordinates of the grid centroid.

SpatialQPFs provides functionality for partitioning the spatial domain into square grids in the `Areal_data()` function. Users can customize the grid size using the `scale` argument, achieving a balance between resolution for detecting local-regional variations and the stability of estimates. Guidelines for optimizing grid size have been previously reported in studies such as Page, D. B. et al.[22], Mi, H. et al.[23], Francis, K. et al.[24], Turkki, R. et al.[25] and Verghese, G. et al.[26].

Within each lattice grid, measurements related to the counts or percentages of each cell type are computed independently, resulting in a mosaic of estimations covering the entire tissue space. Autocorrelation analyses are then conducted to assess the degree of similarity between neighboring locations within the tissue sample for each distinct cell type. These metrics include global and local Moran's I, Geary's C, Getis-Ord statistics, and Lee's L statistics[9,27–31]. Such assessments help identify spatial clustering or dispersion patterns, revealing significant hotspots or coldspots in the distribution of cell types within the tissue, aiding in the identification of regions of elevated or reduced biological activity. Additionally, the degree of colocalization between different cell types is evaluated using methods such as the Morisita-Horn index, Sorensen index, and Jaccard index[32–34]. Features derived from all the aforementioned methods are detailed in Supplementary material Section 1.

Geostatistical data

Geostatistics, a discipline comprised of mathematical and statistical methods, was originally devised to forecast probability distributions of spatial stochastic processes for mining operations. However, its utility has since expanded across a wide range of disciplines. It is utilized in fields such as petroleum geology, earth and atmospheric sciences, agriculture, soil science, and environmental exposure assessment[35–42]. In recent years, geostatistical methods have also been employed in digital pathology images[43,44]. For instance, in a study on breast carcinoma, researchers introduced a novel measurement of heterogeneity using geostatistical methods for histopathological grading tasks[45].

To employ geostatistics techniques for assessing the spatial characteristics of cells, the `Geostatistics_data()` function can be utilized to extract pertinent features. This function employs the variogram, a widely-used method for delineating spatial variability, for feature extraction[46]. The Matérn variogram model is integrated into the function for variogram fitting following the computation of the empirical variogram, which captures the variability measure between pairs of cells separated by different distances. This process unfolds in two key steps. Initially, the empirical variogram is computed, representing a discrete function that captures the variability measure between pairs of cells at varying distances. Subsequently, a theoretical variogram is fitted based on the estimated empirical variogram. From this fitted model, parameters characterizing the spatial variability of cells can be derived. Details regarding the geostatistics features extracted are provided in Supplementary material Section 1.

Example workflow

In this section, we provide brief descriptions of the *SpatialQPFs* functionality.

For first-time use, installation can be completed in the user's local computing environment.

```
devtools::install("your_path/SpatialQPFs", dependencies = TRUE,
repos="https://cloud.r-project.org/")
```

Utilization of *SpatialQPFs* functionality begins with importing input data, followed by preliminary steps such as identifying tumor ROIs, segmenting and classifying cells, before invoking *SpatialQPFs* (Figure 1).

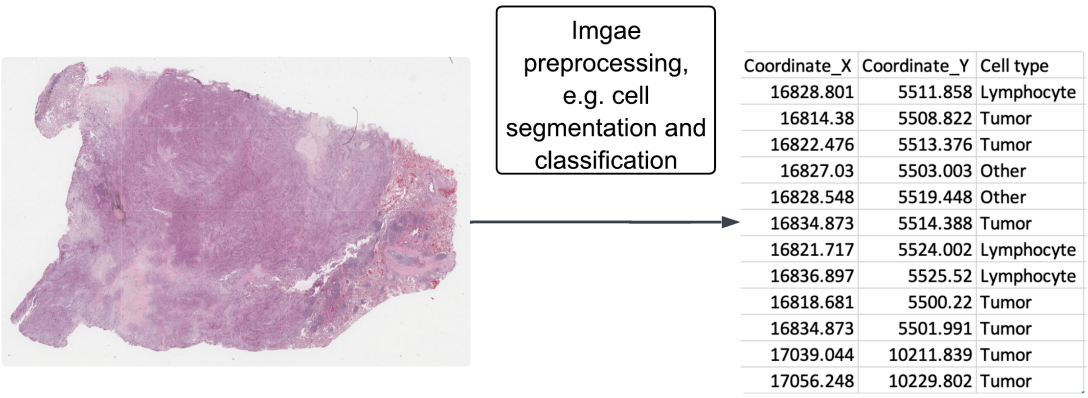

Figure 1. *SpatialQPFs* takes structured tabular cell-level information, such as the centroids of cells and their cell types, as input after performing image preprocessing steps.

To visualize the input tabular data in original tissue space and its derived spatial density map, users can call `Data_Vis()` function. For example, by specifying `cell_type = "Lymphocyte"` users are able to plot the lymphocyte population (Figure 2). See function manual for further details.

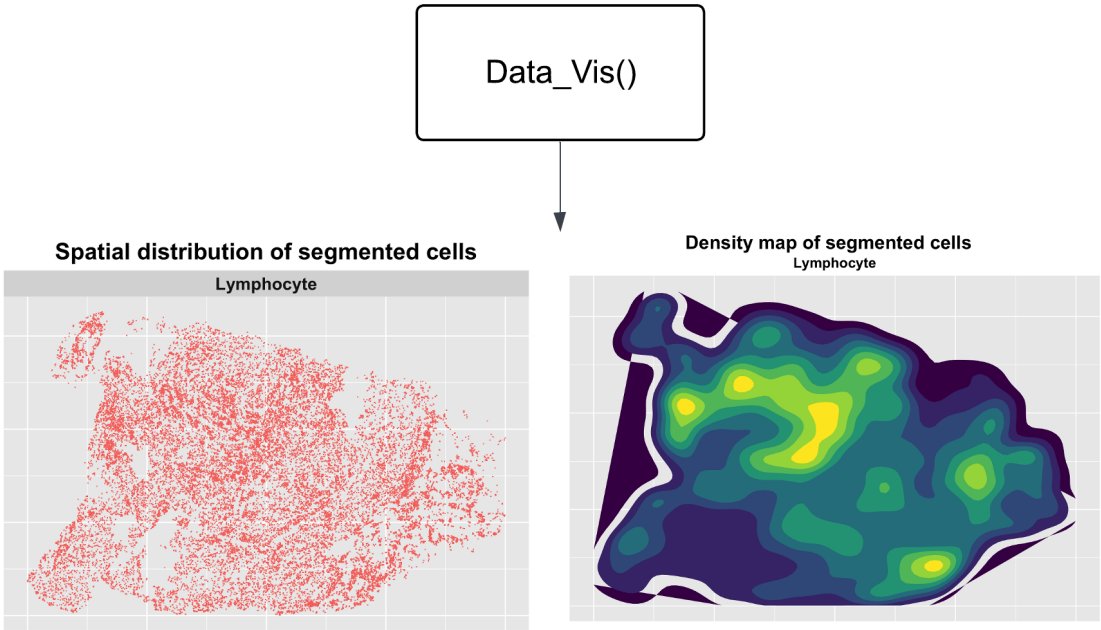

Figure 2. Spatial distribution of user specified cell type and its derived spatial density map.

```
Data_Vis(path = folder_path, file = file_name, cell_class_var = "cell_class",
         x_var = "X0", y_var = "X1", cell_type = "Lymphocyte")
```

The analysis of such structured cellular-level data often begins with point pattern analysis. Point pattern data analysis methods can be conducted by implementing three functions:

- `Point_pattern_data_uni()`: This function generates spatial features for a single cell type population.
- `Point_pattern_data_bi()`: It computes the spatial interaction for a pair of cell types.

- `Point_pattern_data_ITLR()`: This function discriminates a target cell population into subgroups based on their spatial relationships with the reference cell population.

Figure 3 shows selected plots from point pattern analysis, including Multi-type G Function (top left) of Lymphocyte and Tumor cells, Ripley's K Function for Tumor cells (top right) and the localization of Lymphocyte subpopulations (bottom). Additional plots for other point pattern methods are presented at Supplementary material Section 4. Instructions of function arguments can be found in the function manuals.

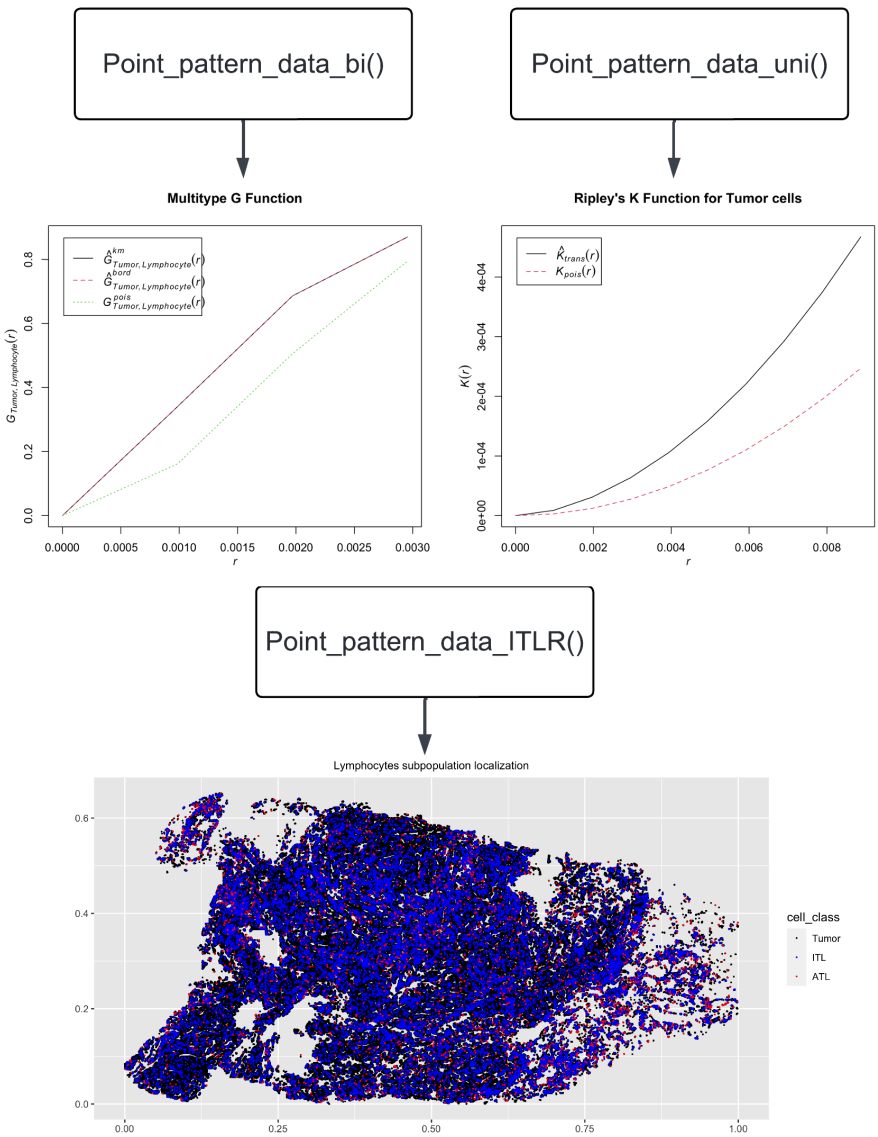

Figure 3. Selected plots from Point pattern analysis, including Multi-type G Function (top left) of Lymphocyte and Tumor cells, Ripley's K Function for Tumor cells (top right) and the localization of Lymphocyte subpopulations (bottom).

```
Point_pattern_data_uni(path = folder_path, file = file_name,
                       cell_class_var = "cell_class", x_var = "X0", y_var = "X1",
                       cell_type = "Lymphocyte", scale = 200, myplot = T)
```

```
Point_pattern_data_bi(path = folder_path, file = file_name,
                      cell_class_var = "cell_class", x_var = "X0", y_var = "X1",
                      from_type = "Lymphocyte",
                      to_type = "Tumor",
                      scale = 200, myplot = T)

Point_pattern_data_ITLR(path = folder_path, file = file_name,
                        cell_class_var = "cell_class", x_var = "X0", y_var = "X1",
                        from_type = "Lymphocyte",
                        to_type = "Tumor",
                        micron_per_pixel = 0.5,
                        myplot = T)
```

Additionally, the analysis of areal data can be performed using the `Areal_data()` function. Note that, this function conducts the transformation from point pattern data to areal data behind the scenes, by partitioning the underlying space into square lattices.

In Figure 4, Local Moran's I map are generated for Tumor cells, wherein the spatial lattices are categorized into "High-High", "High-Low", "Low-High", "Low-Low", and "non-significant" groups according to the prevalence of tumor cells within them and those of neighboring lattices, relative to the mean tumor cell prevalence presented in the sample.

Other areal data analysis plots can be found in Supplementary material Section 4. Detailed instructions for function arguments are available in the function manuals.

```
Areal_data(path = folder_path, file = file_name,
           cell_class_var = "cell_class", x_var = "X0", y_var = "X1",
           from_type = "Lymphocyte",
           to_type = "Tumor",
           scale = 200,
           myplot = T)
```

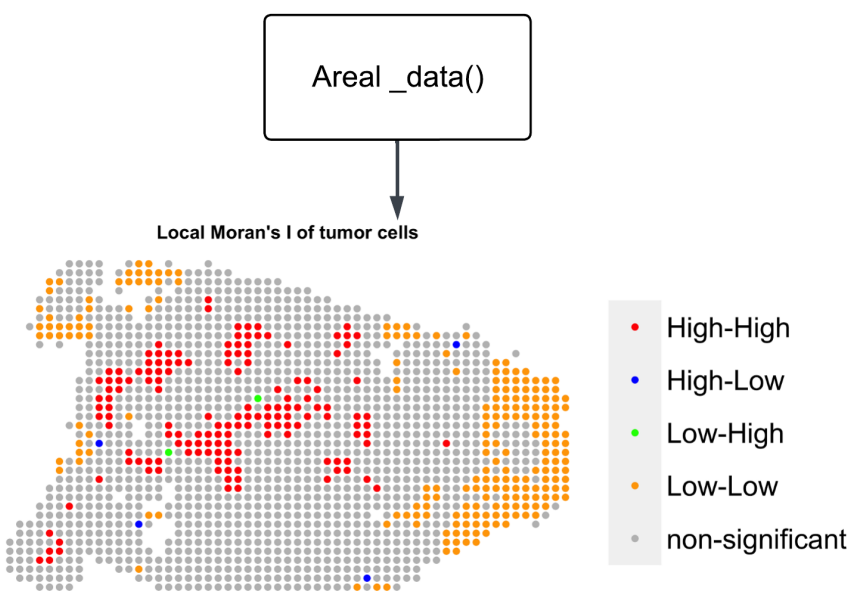

Figure 4. Local Moran's I map for Tumor cells, the spatial lattices are grouped to "High-High", "High-Low", "Low-High", "Low-Low" and "non-significant" according to the prevalence of tumor cells within them and those of neighboring lattices, relative to the mean tumor cell prevalence.

Finally, the analysis of geostatistical data can be conducted using the `Geostatistics_data()` function. In the current version of *SpatialQPFs*, both semi-variogram[46] and cross-variogram[46] are implemented to illustrate the spatial correlation of either a single cell type or a pair of different cell types, respectively.

As depicted in Figure 5, semi-variogram analysis of Tumor cells are performed. Firstly, the empirical semi-variogram is calculated for tumor cell pairs that are separated by various distances. Following this, a smooth curve overlay represents the fitted model that captures the spatial correlation behavior.

Further plots for alternative geostatistical data analysis methods are demonstrated at Supplementary material Section 4. For detailed guidance on function arguments, refer to the function manuals.

```
Geostatistics_data(path = folder_path, file = file_name,
                   cell_class_var = "cell_class", x_var = "X0", y_var = "X1",
                   from_type = "Lymphocyte",
                   to_type = "Tumor",
                   scale = 200,
                   myplot = T)
```

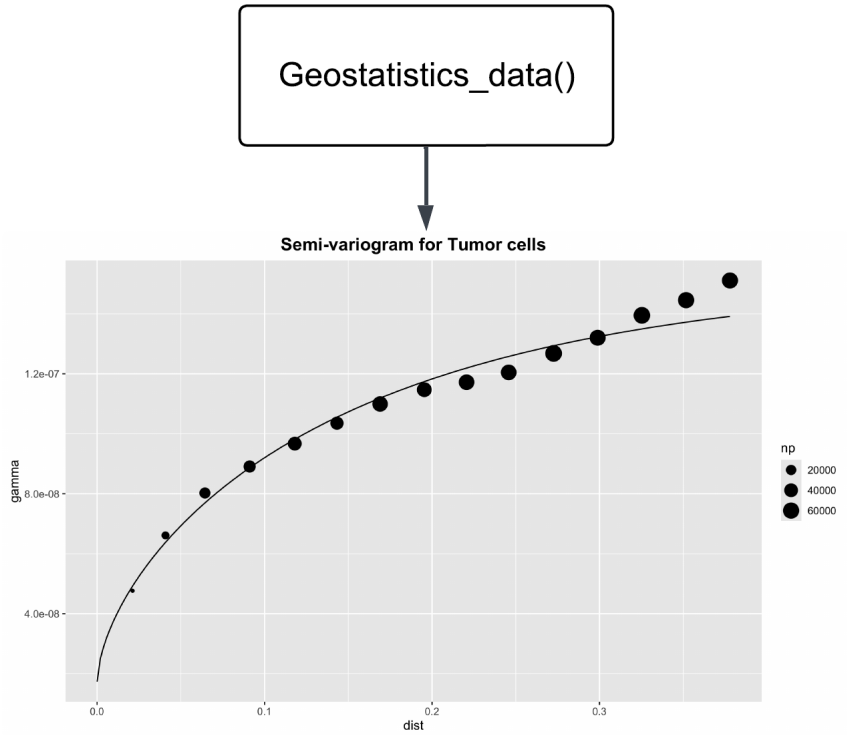

Figure 5. Semi-variogram analysis of Tumor cells.

Additional R output using the example file can be accessed in Supplementary material Section 2.

Applying SpatialQPFs to predict clinical benefit for cancer immunotherapy

To showcase the practical utility of *SpatialQPFs*, we utilized its capabilities to extract spatial features describing the spatial relationship between tumor cells and lymphocytes from all the available H&E stained digital pathology slides. These slides were sourced from two cancer immunotherapy (atezolizumab) clinical trials, IMPower150 (NCT02366143) and IMPower130 (NCT02367781), sponsored by F. Hoffmann–La Roche/Genentech (for details, refer to Supplementary material Section 3). Subsequently, we explored the association between these features and patients' overall survival.

To construct the prediction model incorporating relevant spatial features, we initially, in the training set, employed a univariate Cox proportional-hazards model[47] to select features with a significance threshold of $p \leq 0.15$[48]. Subsequently, we implemented L2-regularized Cox proportional-hazards regression[49] using the selected subset of spatial features to predict patients' survival risks. We designated the ACP arm (atezolizumab plus carboplatin plus paclitaxel) of IMPower150 as our

training set and the ABCP arm (atezolizumab plus carboplatin plus paclitaxel plus bevacizumab) as test cohort 1, for the intra-trial experiment. Additionally, we conducted a cross-trial experiment, utilizing the immunotherapy arm from IMPower130 as test cohort 2.

Figure 6 shows the Kaplan-Meier survival curves of the identified low- and high- risk groups for both experiments. Both experimental results demonstrated that the prediction model based on the derived spatial features effectively stratifies patients into low- and high-risk groups with statistical significance (HR = 0.699 with logrank p = 0.0097 for intra-trial experiment, HR = 0.746 with logrank p = 0.019 for cross-trial experiment). This highlights the predictive power of the extracted features for unseen patients undergoing the same or similar treatment regimens.

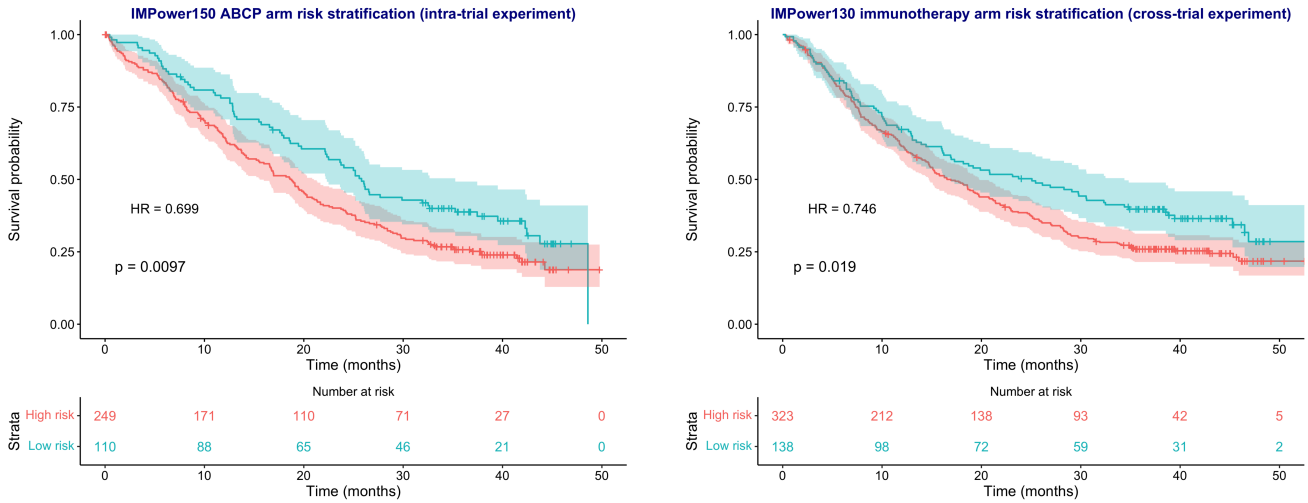

Figure 6. The Kaplan-Meier survival curves of the identified low- and high- risk groups for both experiments. The left panel illustrates the distinct survival risk groups significantly stratified within the intra-trial setting, the right panel displays significantly stratified survival risk groups from the cross-trial setting.

Conclusion and Discussion

We have developed the R package *SpatialQPFs*, tailored for extracting features delineating cell-cell spatial relationships from preprocessed digital pathology data. To the best of our knowledge, this is the first R package that comprehensively implements various spatial statistics methods and derives features based on cell coordinates and cell identities, comprehensively from point process data, areal data, and geostatistical data point of view.

To validate its capabilities, we employed *SpatialQPFs* to extract features that quantify spatial interactions between tumor cells and lymphocytes in two distinct clinical trials. Subsequently, we investigated its capacity to predict patients' survival risks, thereby showcasing its clinical relevance in discerning sub-populations that might derive greater benefits from cancer immunotherapy. This case study not only demonstrates the efficacy of *SpatialQPFs* in extracting meaningful insights from spatial data but also highlights its potential as a valuable tool in clinical research.

It's worth noting that the functionalities of *SpatialQPFs* extend beyond analyzing H&E stained digital images and encompass the analysis of Immunohistochemistry (IHC), Immunofluorescence (IF) staining, as well as spatially resolved single-cell omics data – following the identification of cell types using -omics measurements, the *SpatialQPFs* package can be utilized to investigate cell type interactions. A recent study on IHC image analysis demonstrating this versatility can be found in Li, X. et al.[50].

Several limitations are worth noting. Not all researchers or individuals are proficient in R. With the development of artificial intelligence, Python[51] has become arguably the most popular programming language. To broaden the adoption of the *SpatialQPFs* package, our strategy involves creating additional tutorials and vignettes accessible through Python wrapper functions. One promising solution for this endeavor might be *rpy*2, which serves as an interface to R language within a Python process. Future work will involve leveraging *rpy*2 to expand the accessibility of the package to a wider audience. Furthermore, since image preprocessing serves as a prerequisite steps before invoking *SpatialQPFs*, it is essential to ensure that these preparatory procedures achieve satisfactory performance. For example, the performance of cell segmentation and classification must meet certain standards.

We believe that by leveraging computational tools like *SpatialQPFs*, researchers can effectively harness the vast potential embedded within pathology images, transforming them into a comprehensive array of human-interpretable and mineable

variables on a large scale. This not only streamlines the process of analyzing vast amounts of data but also enables the identification of intricate patterns embedded within the tissue microenvironment. Consequently, this approach accelerates the discovery of novel imaging biomarkers that hold relevance for understanding and diagnosing various human diseases. Through the systematic analysis facilitated by $SpatialQPFs$, researchers gain deeper insights into the complex interplay between cellular structures, paving the way for advancements in disease detection, prognosis, and treatment strategies.

Data availability

For up-to-date details on Roche's Global Policy on the Sharing of Clinical Information and how to request access to related clinical study documents, see here: https://go.roche.com/data_sharing.

References

1. Saltz, J. *et al.* Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell reports* **23**, 181–193 (2018).
2. Hagos, Y. B. *et al.* High inter-follicular spatial co-localization of cd8+ foxp3+ with cd4+ cd8+ cells predicts favorable outcome in follicular lymphoma. *Hematol. Oncol.* **40**, 541–553 (2022).
3. AbdulJabbar, K. *et al.* Geospatial immune variability illuminates differential evolution of lung adenocarcinoma. *Nat. medicine* **26**, 1054–1062 (2020).
4. Sobhani, F. *et al.* Spatial interplay of tissue hypoxia and t-cell regulation in ductal carcinoma in situ. *Npj Breast Cancer* **8**, 105 (2022).
5. Corredor, G. *et al.* Spatial architecture and arrangement of tumor-infiltrating lymphocytes for predicting likelihood of recurrence in early-stage non–small cell lung cancer. *Clin. cancer research* **25**, 1526–1534 (2019).
6. Wang, X. *et al.* Spatial interplay patterns of cancer nuclei and tumor-infiltrating lymphocytes (tils) predict clinical benefit for immune checkpoint inhibitors. *Sci. Adv.* **8**, eabn3966 (2022).
7. Corredor, G. *et al.* Computerized features of spatial interplay of tumor-infiltrating lymphocytes predict disease recurrence in p16+ oropharyngeal squamous cell carcinoma: A multisite validation study. (2020).
8. Nishida-Aoki, N. & Gujral, T. S. Emerging approaches to study cell–cell interactions in tumor microenvironment. *Oncotarget* **10**, 785 (2019).
9. Nawaz, S., Heindl, A., Koelble, K. & Yuan, Y. Beyond immune density: critical role of spatial heterogeneity in estrogen receptor-negative breast cancer. *Mod. Pathol.* **28**, 766–777 (2015).
10. Barua, S. *et al.* Spatial interaction of tumor cells and regulatory t cells correlates with survival in non-small cell lung cancer. *Lung Cancer* **117**, 73–79 (2018).
11. Yuan, Y. Modelling the spatial heterogeneity and molecular correlates of lymphocytic infiltration in triple-negative breast cancer. *J. The Royal Soc. Interface* **12**, 20141153 (2015).
12. Bull, J. A. *et al.* Combining multiple spatial statistics enhances the description of immune cell localisation within tumours. *Sci. reports* **10**, 18624 (2020).
13. Kendall, T. J., Duff, C. M., Thomson, A. M. & Iredale, J. P. Integration of geoscience frameworks into digital pathology analysis permits quantification of microarchitectural relationships in histological landscapes. *Sci. Reports* **10**, 17572 (2020).
14. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2021).
15. Armstrong, R. A. Measuring the spatial arrangement patterns of pathological lesions in histological sections of brain tissue. *Folia Neuropathol.* **44**, 229–237 (2006).
16. Xin, B. *et al.* Dynamic topology analysis for spatial patterns of multifocal lesions on mri. *Med. Image Analysis* **76**, 102267 (2022).
17. Baddeley, A., Rubak, E. & Turner, R. *Spatial point patterns: methodology and applications with R* (CRC Press, 2015).
18. Clark, P. J. & Evans, F. C. Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology* **35**, 445–453 (1954).
19. Yuan, Y. *et al.* Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Sci. translational medicine* **4**, 157ra143–157ra143 (2012).

20. Stoyan, D. & Stoyan, H. *Fractals, random shapes, and point fields: methods of geometrical statistics*, vol. 302 (John Wiley & Sons Inc, 1994).

21. Marcon, E. & Puech, F. Measures of the geographic concentration of industries: improving distance-based methods. *J. Econ. Geogr.* **10**, 745–762 (2010).

22. Page, D. B. *et al.* Spatial analyses of immune cell infiltration in cancer: current methods and future directions: A report of the international immuno-oncology biomarker working group on breast cancer. *The J. pathology* **260**, 514–532 (2023).

23. Mi, H. *et al.* Digital pathology analysis quantifies spatial heterogeneity of cd3, cd4, cd8, cd20, and foxp3 immune markers in triple-negative breast cancer. *Front. physiology* **11**, 583333 (2020).

24. Francis, K. & Palsson, B. O. Effective intercellular communication distances are determined by the relative time constants for cyto/chemokine secretion and diffusion. *Proc. Natl. Acad. Sci.* **94**, 12258–12262 (1997).

25. Turkki, R., Linder, N., Kovanen, P. E., Pellinen, T. & Lundin, J. Antibody-supervised deep learning for quantification of tumor-infiltrating immune cells in hematoxylin and eosin stained breast cancer samples. *J. pathology informatics* **7**, 38 (2016).

26. Verghese, G. *et al.* Multiscale deep learning framework captures systemic immune features in lymph nodes predictive of triple negative breast cancer outcome in large-scale studies. *The J. Pathol.* **260**, 376–389 (2023).

27. Anselin, L. Local indicators of spatial association—lisa. *Geogr. analysis* **27**, 93–115 (1995).

28. Besag, J. Spatial interaction and the statistical analysis of lattice systems. *J. Royal Stat. Soc. Ser. B (Methodological)* **36**, 192–225 (1974).

29. Moran, P. A. Notes on continuous stochastic phenomena. *Biometrika* **37**, 17–23 (1950).

30. Geary, R. C. The contiguity ratio and statistical mapping. *The incorporated statistician* **5**, 115–146 (1954).

31. Lee, S.-I. A generalized significance testing method for global measures of spatial association: an extension of the mantel test. *Environ. Plan. A* **36**, 1687–1703 (2004).

32. Maley, C. C., Koelble, K., Natrajan, R., Aktipis, A. & Yuan, Y. An ecological measure of immune-cancer colocalization as a prognostic factor for breast cancer. *Breast Cancer Res.* **17**, 131 (2015).

33. Rempala, G. A. & Seweryn, M. Methods for diversity and overlap analysis in t-cell receptor populations. *J. mathematical biology* **67**, 1339–1368 (2013).

34. Chao, A., Chazdon, R. L., Colwell, R. K. & Shen, T.-J. A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecol. letters* **8**, 148–159 (2005).

35. Chien, L.-C., Guo, Y., Li, X. & Yu, H.-L. Considering spatial heterogeneity in the distributed lag non-linear model when analyzing spatiotemporal data. *J. exposure science & environmental epidemiology* **28**, 13–20 (2018).

36. Chien, L.-C., Li, X. & Staudt, A. Physical inactivity displays a mediator role in the association of diabetes and poverty: A spatiotemporal analysis. *Geospatial health* (2017).

37. Li, X., Staudt, A. & Chien, L.-C. Identifying counties vulnerable to diabetes from obesity prevalence in the united states: a spatiotemporal analysis. *Geospatial health* (2016).

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

39. Liu, Y. *et al.* Spatial epidemiology and spatial ecology study of worldwide drug-resistant tuberculosis. *Int. journal health geographics* **10**, 50 (2011).

40. Li, X., Guindani, M., Ng, C. S. & Hobbs, B. P. A bayesian nonparametric model for textural pattern heterogeneity. *J. Royal Stat. Soc. Ser. C: Appl. Stat.* **70**, 459–480 (2021).

41. Li, X., Guindani, M., Ng, C. S. & Hobbs, B. P. Spatial bayesian modeling of glcm with application to malignant lesion characterization. *J. applied statistics* **46**, 230–246 (2019).

42. Li, X., Guindani, M., Ng, C. S. & Hobbs, B. P. Classification of adrenal lesions through spatial bayesian modeling of glcm. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, 147–151 (IEEE, 2017).

43. Li, Q., Wang, X., Liang, F. & Xiao, G. A bayesian mark interaction model for analysis of tumor pathology images. *The annals applied statistics* **13**, 1708 (2019).

44. Santos, J. D., Silva, R. R. & Veras, R. M. Application of geostatistical functions and deep features to kidney biopsy images to differentiate focal segmental glomerulosclerosis from minimal change disease. In *Anais Estendidos do XXXIII Conference on Graphics, Patterns and Images*, 56–62 (SBC, 2020).

45. Sharifi-Salamatian, V., de Roquancourt, A. & Rigaut, J. P. Breast carcinoma, intratumour heterogeneity and histological grading, using geostatistics. *Anal. Cell. Pathol.* **20**, 83–91 (2000).

46. Cressie, N. Statistics for spatial data. *Terra Nova* **4**, 613–617 (1992).

47. Cox, D. R. Regression models and life-tables. *J. Royal Stat. Soc. Ser. B (Methodological)* **34**, 187–202 (1972).

48. Chowdhury, M. Z. I. & Turin, T. C. Variable selection strategies and its importance in clinical prediction modelling. *Fam. medicine community health* **8** (2020).

49. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for cox's proportional hazards model via coordinate descent. *J. statistical software* **39**, 1 (2011).

50. Li, X. *et al.* Automated tumor immunophenotyping predicts clinical benefit from anti-pd-l1 immunotherapy. *J. Pathol.* **263**, 190–202 (2024).

51. Van Rossum, G., Drake, F. L. *et al. Python reference manual*, vol. 111 (Centrum voor Wiskunde en Informatica Amsterdam, 1995).

Acknowledgements

The author expresses gratitude to all members of phc-DP, Genentech, Inc. / F. Hoffmann-La Roche AG and Computational Science and Informatics, Roche Diagnostic Solution, Pathology Lab for their invaluable feedback and assistance in curating the clinical images and outcome data. The clinical trials received support from F. Hoffmann-La Roche Ltd./Genentech, Inc. Additionally, the author extends heartfelt thanks to the patients and their families.

Author contributions statement

XL conceived and designed the study, conducted the experiments, and analysed the results.

Competing interests

XL is a full-time employee of Genentech, Inc./Roche and owns stock from Roche.

Additional information

Supplementary Information The Supplementary material is available at online version of the journal.