PLOS COMPUTATIONAL BIOLOGY

METHODS

Using random forests to uncover the predictive power of distance-varying cell interactions in tumor microenvironments

Jeremy VanderDoes[1], Claire Marceaux[2,3], Kenta Yokote[2], Marie-Liesse Asselin-Labat[2,3], Gregory Rice[1], Jack D. Hywood[4]*

1 Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada, **2** Personalised Oncology Division, The Walter and Eliza Hall Institute of Medical Research, Parkville, Australia, **3** Department of Medical Biology, The University of Melbourne, Parkville, Australia, **4** Department of Anatomical Pathology, Royal Melbourne Hospital, Melbourne, Australia

* jack.d.hywood@gmail.com

Citation: VanderDoes J, Marceaux C, Yokote K, Asselin-Labat M-L, Rice G, Hywood JD (2024) Using random forests to uncover the predictive power of distance-varying cell interactions in tumor microenvironments. PLoS Comput Biol 20(6): e1011361. https://doi.org/10.1371/journal.pcbi.1011361

Editor: Pedro Mendes, University of Connecticut School of Medicine, UNITED STATES

Received: July 18, 2023

Accepted: May 31, 2024

Published: June 14, 2024

Data Availability Statement: The data and source code used to produce the results and analyses presented in this manuscript are available from the Supporting information files and a GitHub Repository (https://github.com/jrvanderdoes/funkycells). The code is also explained on the package website, https://jrvanderdoes.github.io/funkycells. Data is also provided in the supporting information (S1 Table).

Abstract

Tumor microenvironments (TMEs) contain vast amounts of information on patient's cancer through their cellular composition and the spatial distribution of tumor cells and immune cell populations. Exploring variations in TMEs between patient groups, as well as determining the extent to which this information can predict outcomes such as patient survival or treatment success with emerging immunotherapies, is of great interest. Moreover, in the face of a large number of cell interactions to consider, we often wish to identify specific interactions that are useful in making such predictions. We present an approach to achieve these goals based on summarizing spatial relationships in the TME using spatial *K* functions, and then applying functional data analysis and random forest models to both predict outcomes of interest and identify important spatial relationships. This approach is shown to be effective in simulation experiments at both identifying important spatial interactions while also controlling the false discovery rate. We further used the proposed approach to interrogate two real data sets of Multiplexed Ion Beam Images of TMEs in triple negative breast cancer and lung cancer patients. The methods proposed are publicly available in a companion `R` package `funkycells`.

Author summary

Spatial data on the tumor microenvironment (TME) are becoming more prevalent. Existing methods to interrogate such data often have several limitations: (1) they can rely on estimating the spatial relationships among cells by examining simple counts of cells within a *single* radius, (2) they may not come with ways to evaluate the statistical significance of any findings, or (3) they model individual interactions independently of other interactions. Our approach leverages techniques in spatial statistics and uses a benchmark ensemble machine learning method to address each of these deficiencies; it (1) uses *K* functions to encode the relative densities of cells over all radii up to a user-selected

Funding: M.L.A.L. is supported by funding from the Viertel Foundation Senior Medical Research Fellowship and the 'ACRF Program for resolving cancer complexity and therapeutic resistance', proudly supported by Australian Cancer Research Foundation. G.R. was supported by the Natural Science and Engineering Research Council of Canada, Discover Grant RGPIN-03723. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

maximum radius, (2) employs permutation and cross-validation to evaluate the statistical significance of any findings on the spatial interactions in the TME, and (3) models multiple interactions simultaneously. Our approach is freely available with an `R` implementation called `funkycells`. In the analysis of two real data sets, we have seen that the method performs well, and gives the expected results. We think this will be a robust tool for researchers looking to interrogate TME data.

Introduction

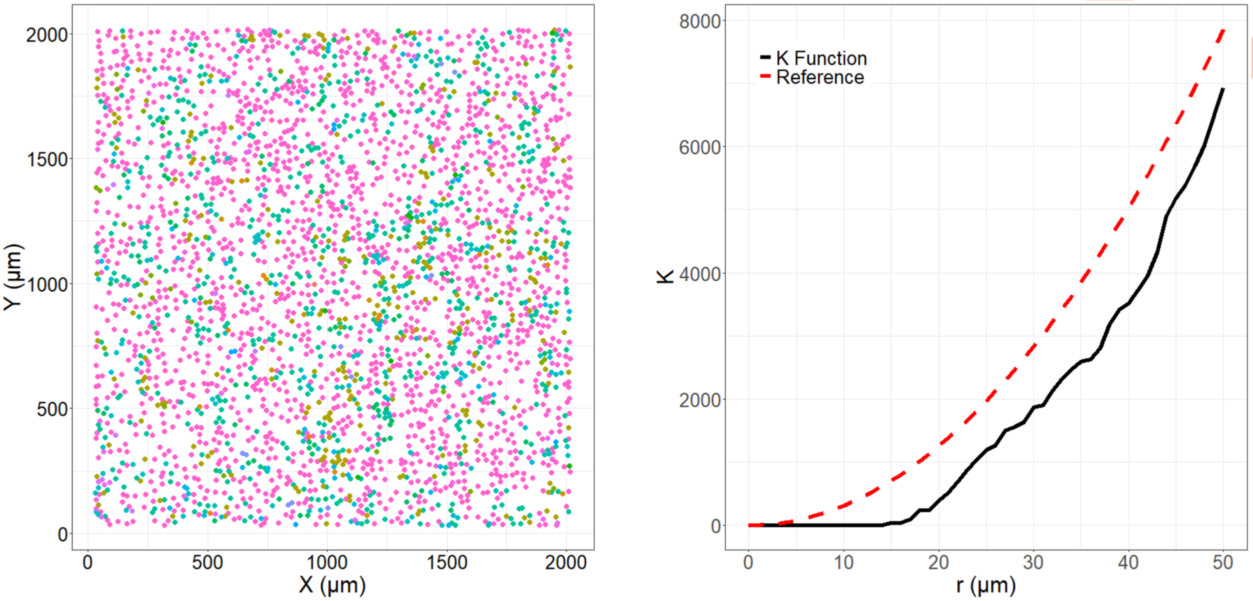
Recent advances in cancer treatment, such as immune checkpoint inhibition and other cancer immunotherapies, have sparked a growing interest in studying the cellular composition and spatial organization of the tumor microenvironment (TME). The latest innovations in imaging technologies allow for single cell resolution of specific proteins, facilitating in-depth study of the spatial arrangement of cell types within the TME. A wide variety of technologies are available for this purpose, each with different benefits and trade-offs [1–7]. For a review of the available technologies see [8].

In comparing TME data, different spatial relationships between cell types, e.g. between tumor cells and specific immune cell populations, and/or individual proteins, often appear predictive of patient outcomes and may guide therapeutic interventions; see for example [9]. Comparisons between cancer subtypes, e.g. hormone-positive versus hormone-negative breast cancers, or lung squamous cell carcinoma vs lung adenocarcinoma, may provide novel insight into tumor biology and guide the development of treatments. A further goal is to identify specific spatial relationships observed in particular patient's tumor that are useful in predicting an outcome, such as patient survival or response to therapy. Recent results demonstrate that TME data can be used for such prediction in a variety of tumor types [10, 11].

We consider such prediction problems for data sets generated from tumors imaged with Multiplexed Ion Beam Imaging (MIBI) by means of the MIBIscope in this paper. The MIBIscope uses ion-beam ablation and time-of-flight mass spectrometry to detect up to approximately 40 protein markers on formalin-fixed, paraffin-embedded (FFPE) tissue. Thus, it provides comprehensive data on cell characteristics and their localisation at a single-cell resolution of around 250–400nm [12, 13]. Such data collected on the TME can be considered as marked spatial point patterns [14–16]. The cell locations can be considered as points within the pattern, with cell phenotypes and/or protein markers giving the "marks". An example of such a point process generated from a tumor imaged using MIBIscope is shown in Fig 1a [12].

Existing methods developed to study cellular interactions in the TME have exploited cell neighborhood analysis in which the spatial relationship between a cell of interest and its neighboring cells can reveal particular cell-cell interactions associated with a disease state or changes associated with response to therapy; see e.g. [17]. Pairwise cell-to-cell distance calculations over iterations of randomized permutations has also been used to identify relevant cell-cell interactions [18]. However, the substantial number of cell types present in the TME leads to a very large number of potential pairwise interactions creating a major challenge in finding interactions that may be meaningful and statistically significant in predicting outcomes of interest. There is ongoing interest in applying spatial statistics methods to similar biological data sets, e.g. [19–21].

A common method of analysing spatial point patterns, such as those that arise in TME imaging, is to consider Ripley's K function [22–24]. The K function describes the distribution of inter-point distances in a given point pattern, giving an indication as to whether points in



(a) Point pattern

(b) K function

Fig 1. An example of point pattern data and an associated *K* function. (a) Point pattern data associated with a tumor imaged using MIBIscope from a triple negative breast cancer patient with multiple identified phenotypes. The *x*- and *y*-axes represent the spatial dimensions, with the points giving individual cell locations, and the colour of the points indicating one of 15 unique phenotypes, e.g. tumor (red), NK cells (purple), and monocytes/neutrophils (cyan). (b) The associated cross *K* function (black) for two cell types in the image: tumor and monocytes/neutrophils. The *x*-axis indicates the radius, *r*, and the *y*-axis gives the value of the *K* function. The estimated *K* function can be compared to πr^2 (red dashed line), which is the theoretical *K* function associated with complete spatial randomness.

https://doi.org/10.1371/journal.pcbi.1011361.g001

the pattern (e.g. cells) are clustered or dispersed with respect to one another. The *K* function, along with other summary functions from spatial statistics, has previously been employed in the analysis of the TME [25–31]. An example of a *K* function showing the relative distribution of a specific immune cell type around tumor cells within a MIBIscope image from a triple negative breast cancer patient is shown in Fig 1b.

In this paper, we present a general framework for analyzing and identifying useful spatial relationships in the TME through predicting an outcome of interest. The method we propose uses a novel combination of spatial statistics and functional data analysis, in conjunction with methods in ensemble machine learning. The application of functional data analysis to spatial point pattern data is a recent development [31–35].

Our approach begins by producing *K* functions for the different cell-cell, (or alternatively marker-marker), interactions within images. After performing dimension reduction using functional principal components analysis [36], these data are combined with non-functional patient meta-data, such as age or sex, and a modified random forest model is used to predict the patient outcome. Motivated by [37], in order to evaluate the predictive power of the spatial interactions, "knock-off" point patterns that mimic the spatial data in the TME are generated, via permutation, independently of the responses. The importance of specific spatial interactions in predicting the response are evaluated by comparison to the predictive power of the knock-off spatial patterns. This approach overcomes the challenge of distinguishing important spatial interactions among many potential interactions of interest, i.e. it controls the false discovery rate. It also lends itself to the generation of easy-to-interpret plots showing which interactions are useful in predicting the response. Moreover, it grants high power for even a relatively few number of cells due to the robustness of the *K* functions, and high power even with small sample sizes. Small sample sizes are a relatively common obstacle for this type of data.

To our knowledge the method we propose is the first to evaluate for the importance of spatial interactions in the TME that may vary over distance with a large number of potential cell interactions. Many statistical methods proposed to date rely on simple summaries of the images, such as the average number of cells observed, or otherwise simplify the data by, for example, initially clustering cells into groups (e.g. [10, 18, 38–41]). In other comparable methods available in the literature, a single distance of interest is considered (e.g. [12, 26, 42]). [30] provides an approach for detecting differences across multiple images between cell-cell interactions by comparing the integrated difference between a spatial summary function similar to the *K* function and its expected value under complete spatial randomness. However, this approach only considers individual spatial interactions, while we wish to consider all interactions in a single model. Additionally, a possible drawback of the approach in [30] is that integration over the summary statistic may lose valuable information relating to differences in the shape of the functions, since differently shaped spatial summary functions may still have the same integral. [31] uses a related functional linear model based on interaction distance functions, although this approach is not tailored to examine a large number of interactions simultaneously.

Some methods may predict well, but do not lead to interpretable results that allow for the importance of individual spatial characteristics to be compared. Other methods consider only a single image or an equal number of images per patient. This is further complicated by the fact that some images do not contain all cell types. However, our approach is interpretable and able to be used to analyse data with multiple, possibly differing, numbers of observations per patient. Additionally, images may have different shapes or sizes. In that sense, our approach allows for the complete use of the data.

We apply the proposed methods to two MIBIscope data sets; a data set of triple negative breast cancer (TNBC) patients, and a data set consisting of both lung squamous cell carcinomas (LUSC) and lung adenocarcinomas (LUAD). Regarding the TNBC data, our method was accurately able to identify clustered versus dispersed tumors when compared to [12], and was additionally able to identify important cell spatial interactions in making that determination. Our method also indicated that there did not appear to be measurable differences in the spatial arrangement of tumor and immune cell types, as measured by homogeneous K functions, between the LUSC and LUAD groups.

Whilst the methodology presented here is motivated by, and applied to, MIBIscope data, it can be applied to similar data generated by other technologies, e.g. OPAL, Phenocycler Fusion, Merscope, Xenium and Cosmx [43–49]. Furthermore, the methodology can easily extend beyond two-dimensions to higher-dimensional images, another area of active research [50].

The rest of the paper is organized as follows. In the section Materials and methods, we give a detailed description of the data we consider and the methods to analyze them, including subsections on how we fit a modified random forest in this setting, and how we evaluate the statistical significance and uncertainty in measuring the variable importance of spatial interactions of cell types as encoded by *K* functions. We also introduce the `R` package `funkycells` (shortened version of **fun**ctional data analysis of **K** functions for multiplexed images of **cells**), an open-source implementation of our approach in that section. The Simulation study section details the results of simulation experiments in which we found that the proposed method performed well when applied to synthetic data built to mimic the TNBC data. We report the results when this approach was applied to the TNBC and LUSC vs. LUAD data sets in the section Applications to MIBIscope data. Some concluding remarks and directions for future work are collected in Discussion.

Materials and methods

The raw spatial data that we consider take the form of 2-dimensional point patterns, as generated using MIBIscope. We denote the cell spatial data, as

$$\mathbf{C} = \{(x_{c,t}^{(p,i)}, y_{c,t}^{(p,i)}, \mathbf{a}_{c,t}^{(p,i)}), p = 1, \ldots, N,\ i = 1, \ldots, I_p,\ t = 1, ..., T,\ c = 1, \ldots, n_{p,i,t}\}, \quad (1)$$

where $(x_{c,t}^{(p,i)}, y_{c,t}^{(p,i)}, \mathbf{a}_{c,t}^{(p,i)})$ denotes the x and y coordinates and the associated cell properties **a** of a given cell. Specifically, the term refers to the the c^{th} cell of type t (of which there are T total types), for the i^{th} image of the p^{th} patient, with $n_{p,i,t}$ giving the number of cells of type t in image i of patient p. The properties in $\mathbf{a}_{c,t}^{(p,i)}$ may (typically) simply give the cell's phenotype (and is therefore redundant due to the term t), or may be more general, such as a vector describing individual protein expression–allowing use of this method for processed or raw data. For example, the vector could be composed of binary indicators as to whether a protein is expressed or not. For notational clarity, and since we here only consider data consisting of cells and their associated phenotype, we drop the **a** term throughout the paper.

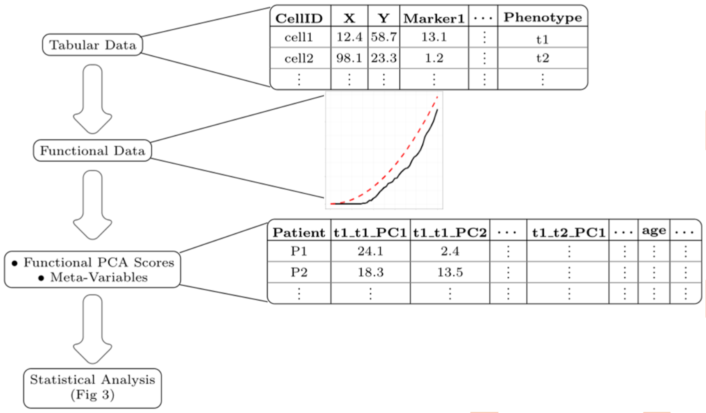
Since the applicability of our method extends beyond this example, we designate several general terms for use throughout the paper. We refer to point patterns such as in Fig 1a as "images". We interchangeably use the terms cell phenotype and cell type. We also interchangeably use the terms cross-over K function and K function. We assume a single response variable, Z_i, for each of the N patients (e.g. tumor type, response to therapy, etc.). The set of outcomes for the N patients is denoted $\mathbf{Z} = (Z_1, \ldots, Z_N)$. In the real data examples we consider below, Z_i is a binary response, e.g. "compartmentalized" versus "mixed" tumors for the TNBC data, or LUSC versus LUAD for the lung cancer data, in which case we can encode the outcomes as taking the values 0 and 1. These methods may easily be adapted for more general class responses, e.g. different types of tumors, or numeric responses, e.g. survival time.

In addition to the spatial data, we assume that we may have access to non-spatial data on the patients. We refer to this data as patient "meta-data", and we assume that it takes the form $\mathbf{M} = (\mathbf{m}_1, \ldots, \mathbf{m}_N)$, where each $\mathbf{m}_i$ is a vector of patient attributes, for example age or sex.

With both the cell spatial data **C** and meta-data **M**, our goals are to (1) investigate to what extent these data are useful in predicting the outcomes **Z**, and (2) to identify which specific spatial relationships and/or components of the meta-data from the full data set are useful in predicting **Z**. We deem data on a spatial relationship or component of the meta-data "useful" if their importance in predicting the outcome exceeds, to a statistically significant degree, that of similar variables that are known to be unrelated to the outcome. For reference throughout the paper, high-level schematics of our proposed method are presented in Figs 2 and 3. Fig 2 overviews the major steps for processing data in our model while Fig 3 focuses on the steps of the statistical analysis of the model.

Towards answering these questions, we build a model of the outcomes **Z** in terms of the image spatial information **C** and meta-data **M**. In doing so, we must address how we incorporate the complex image data into such a model. Motivated by the expectation that patient outcomes are influenced by the relative distribution of various immune cells or protein markers around each other, we begin by computing spatial "cross-over" K functions from the image data, which summarize the spatial distribution of cells with respect to one another as in Fig 1b.

We provide an open source implementation of our approach in R [51] on CRAN, in the package, `funkycells`, or at the site github.com/jrvanderdoes/funkycells. This



CellID	X	Y	Marker1	⋯	Phenotype
cell1	12.4	58.7	13.1	⋮	t1
cell2	98.1	23.3	1.2	⋮	t2
⋮	⋮	⋮	⋮	⋮	⋮

Patient	t1_t1_PC1	t1_t1_PC2	⋯	t1_t2_PC1	⋯	age	⋯
P1	24.1	2.4	⋮	⋮	⋮	⋮	⋮
P2	18.3	13.5	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Fig 2. Flow chart for data processing. The methods presented here begin with tabular data obtained after pre-processing multiplex images (steps that include cell segmentation, phenotyping, etc.). For a given image the tabular data consists of rows for each imaged cell, giving the associated x-y position, marker intensities, and cell phenotype. Next, the tabular data are converted into spatial *K* functions for each interaction of interest (this can be exhaustive, and include all possible interactions between phenotypes, or selective, with only a subset of interactions analysed). Next, *K* functions are converted into functional principal component scores. Patient meta-variables are added at this stage. The resulting data is then used in the statistical model, as described in Fig 3.

https://doi.org/10.1371/journal.pcbi.1011361.g002

implementation is runnable on personal computers, and includes the code and data used in the presented simulations and data analyses.

Summarizing the image data using K functions

The cross-over *K* function for image *i* of person *p* and cell types *t* and *t*′ is defined as

$$K_{t,t'}^{(p,i)}(r) = \frac{1}{\lambda_{t'}} \mathrm{E}(\text{Number of cells of type } t' \text{ within distance } r \text{ of a cell of type } t), \qquad (2)$$

where E denotes mathematical expectation, the radius r ranges from $0 \leq r \leq R$, with the max radius R being a user specified parameter that we discuss below, and $\lambda_{t'}$ gives the density of cells of type t' [14–16].

By examining this function for varying radii, we may infer how cell types are distributed around each other. For example, if cell types are distributed around each other entirely at random, then $K_{t,t'}^{(p,i)}(r)$ is equal to the area of a circle of radius r, πr^2. Regularity or dispersion of the cells around each other tends to reduce $K_{t,t'}^{(p,i)}(r)$ while clustering tends to increase it. An example of a *K* function computed between the tumor and monocytes/neutrophils phenotypes for a given tumor in the TNBC data set is presented in Fig 1b, which indicates a degree of dispersion with respect to monocytes/neutrophils around tumor cells across the given *r* values compared to that expected for cells distributed around tumor cells with complete spatial randomness. Cross-over *K* functions can be used to summarize all two-way interactions between cell types for a given image.

In practice, estimation is based on an empirical average replacing the expectation. The estimated cross-over *K* function for image *i* of person *p* is given by

$$K_{t,t'}^{(p,i)}(r) = \frac{|i|}{n_{p,i,t'}} \sum_{c=1}^{n_{p,i,t}} \sum_{c'=1}^{n_{p,i,t'}} \mathbb{1}\left(\sqrt{[x_{c,t}^{(p,i)} - x_{c',t'}^{(p,i)}]^2 + [y_{c,t}^{(p,i)} - y_{c',t'}^{(p,i)}]^2} \leq r\right), \qquad (3)$$

$0 \leq r \leq R$, where indicator $\mathbb{1}\,(A)$ takes the value one if the condition A is satisfied, and is zero

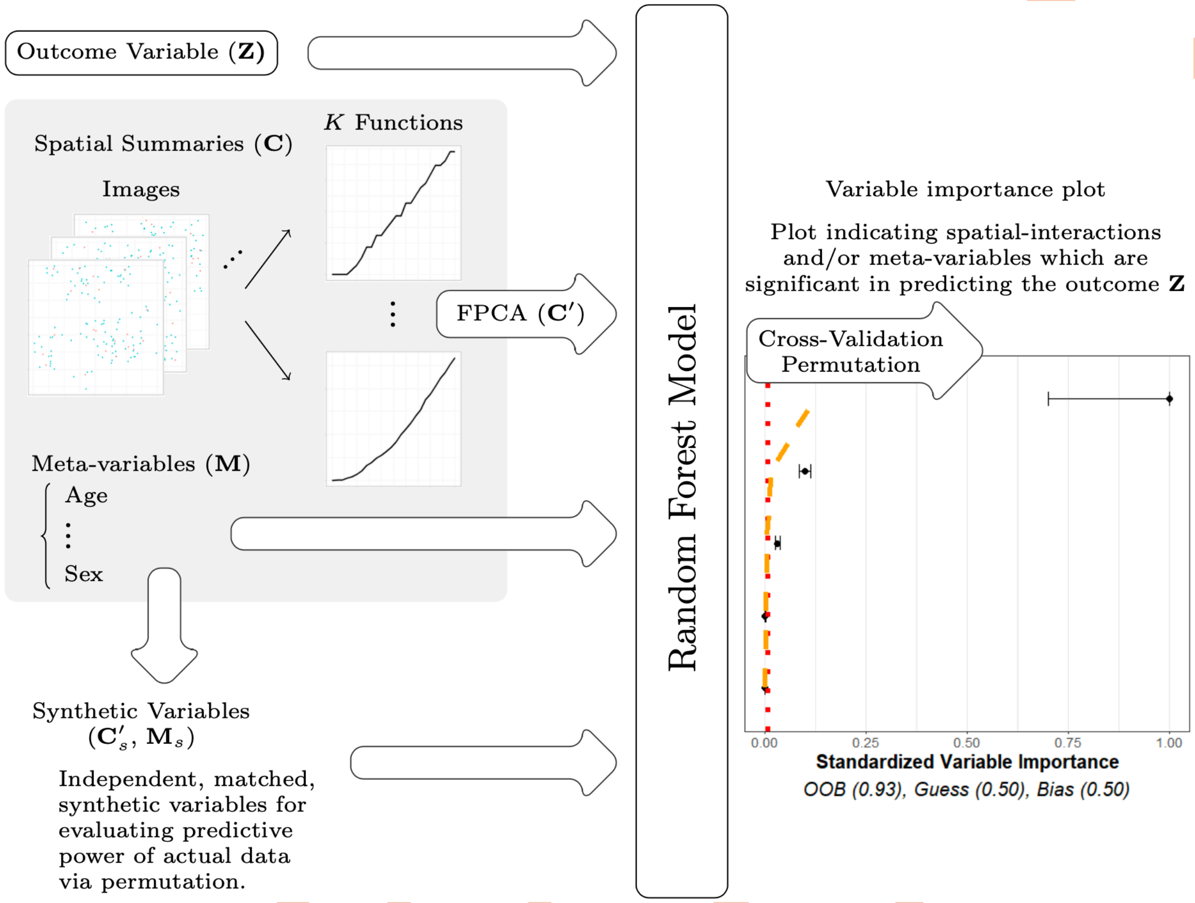

Outcome Variable (**Z**)
Spatial Summaries (**C**)
Images
K Functions
FPCA (**C′**)
Meta-variables (**M**)
Age
Sex
Synthetic Variables
($\mathbf{C}'_s$, $\mathbf{M}_s$)
Independent, matched, synthetic variables for evaluating predictive power of actual data via permutation.
Random Forest Model
Variable importance plot
Plot indicating spatial-interactions and/or meta-variables which are significant in predicting the outcome **Z**
Cross-Validation
Permutation
0.00
0.25
0.50
0.75
1.00
Standardized Variable Importance
OOB (0.93), Guess (0.50), Bias (0.50)

Fig 3. Flow chart of model. When modeling using `funkycells`, there are several major steps: organizing data, generating synthetic data, and modeling using random forests. The spatial data is organized into functional summaries (K functions) that are projected into finite dimensions (FPCA) and used with meta-variables to predict the outcome variable. The spatial data and meta-variables are permuted to create synthetic variables with similar properties but independent of the outcome. These synthetic variables are then added to the model, and used to quantify the strength of the relationships between the spatial and meta-data with the response. The model processes the data, employing cross-validation and permutation to return a variable importance plot (with predictive accuracy estimates) indicating spatial interactions and/or meta-variables which are significant in predicting the outcome **Z**.

https://doi.org/10.1371/journal.pcbi.1011361.g003

otherwise, and $|i|$ indicates the area of the image. When patients have multiple images, we combine their cross-over K functions by computing a weighted average,

$$K^{(p)}_{t,t'}(r) = \sum_{i=1}^{I_p} \frac{n_{p,i,t}}{n_{p,\cdot,t}} K^{(p,i)}_{t,t'}(r), \quad 0 \le r \le R\,, \tag{4}$$

with I_p giving the number of images for a given patient p. In other words, the K functions from each image are weighted according to the prevalence of the cells of the type under consideration. We note that if there is one image per patient (so that $I_p = 1$), then $K^{(p)}_{t,t'}(r) = K^{(p,1)}_{t,t'}(r)$, and further that the weights in Eq (4) vanish if the cell types t and t' are missing in an image. We use a standard isotropic edge correction in this paper (see Appendix A).

In computing these K functions for each cell type, we can transform the spatial data **C** into a collection of T^2 different K functions for each patient, $\{K^{(p)}_{t,t'}(r),\ t, t' = 1, \ldots, T,\ 0 \le r \le R\}$. The K functions are then treated as functional data objects; see e.g. [36]. Since even moderate values of T lead to a large number of K functions to consider, user input is often helpful in determining a smaller subset of interactions (and hence K functions) of particular interest for analysis.

Although informative, these K functions are unwieldy to directly use in a model, and we further transform the functions using the dimension reduction technique of functional principal component analysis (FPCA). FPCA is a common technique in functional data analysis that decomposes the leading sources of variability among the curves $K_{t,t'}^{(p)}(r)$ into a set of finite-dimensional, approximately uncorrelated principal components (PCs); see [36]. To do so, for each pair of cell types t and t', we define the empirical covariance kernel as

$$\hat{C}_{t,t'}(r,s) = \frac{1}{N}\sum_{p=1}^{N}\left[K_{t,t'}^{(p)}(r) - \bar{K}_{t,t'}(r)\right]\left[K_{t,t'}^{(p)}(s) - \bar{K}_{t,t'}(s)\right], \tag{5}$$

where $\bar{K}_{t,t'}(r) = \frac{1}{N}\sum_{p=1}^{N} K_{t,t'}^{(p)}(r)$.

The eigenvalues and eigenfunctions of the kernel $\hat{C}_{t,t'}$ are then computed to satisfy the functional equation

$$\lambda_{i,t,t'}\phi_{i,t,t'}(r) = \int_0^R \hat{C}_{t,t'}(r,s)\phi_{i,t,t'}(s)ds\,. \tag{6}$$

The K function $K_{t,t}^{(p)}$ is summarized using the d coefficients (PCs)

$$\mathbf{k}_{t,t'}^{(p,d)} = \left(\int_0^R K_{t,t}^{(p)}(r)\phi_{1,t,t'}(r)dr, \ldots, \int_0^R K_{t,t}^{(p)}(r)\phi_{d,t,t'}(r)dr\right)^{\top}. \tag{7}$$

The d coefficients comprising $\mathbf{k}_{t,t'}^{(p,d)}$ describe the projection of the K function $K_{t,t}^{(p)}$ onto the finite dimensional linear space spanned by $\phi_{1,t,t'}, \ldots, \phi_{d,t,t'}$, which are optimal in terms of capturing the variability among the curves $K_{t,t}^{(p)}$, $p = 1, \ldots, N$, with a d-dimensional summary. An advantage of summarizing the curves in this way is that, when differences in the K functions across the population are present due to differences in the outcome(s) of interest, the PCs are expected to capture these differences.

As such, we summarize the spatial data using the principal components $\mathbf{C}' = \{\mathbf{k}_{t,t'}^{(p,d)}, p = 1, \ldots, N,\ t, t' = 1, \ldots, T\}$, which we then incorporate with the meta-data $\mathbf{M}$ into a model for $\mathbf{Z}$ of the form

$$\hat{\mathbf{Z}} = f(\mathbf{C}', \mathbf{M}). \tag{8}$$

Since our ultimate goal includes evaluating which spatial interactions or elements of the meta-data are useful in predicting the outcomes, we use a random forest model for f. Random forest models are tree-based ensemble machine learning methods in which decision trees are built, after sampling with replacement the patient data and discarding some covariates at random, by sequentially splitting on variables to minimize a metric for predicting $\mathbf{Z}$ [52]. The main reasons for the sampling procedures for the patient data and covariates in building each tree is to build nearly independent trees and also address overfitting, common in many machine learning applications. When the trees are combined to create a forest, increased statistical power is observed. Additionally, the computational complexity is similar to a traditional random forest model, see Results.

Variable importance

Random forest models are useful in achieving our goals since they have strong predictive power while still allowing for a quantification of the usefulness of individual covariates in predicting the response through various "variable importance" measures. We now describe the computation of variable importance metrics for the random forest models introduced [52]. There are multiple methods for calculating variable importance values, such as permuting variables in data and comparing the difference in loss metrics, or measuring node purity. We tested several methods and found similar results regardless of the variable importance metric used. Due to computational considerations, we implement a node purity metric, sometimes called recursive partitioning, as described in [53, 54].

In the following explanations for node purity variable importance, it is perhaps useful to imagine the scenario where the outcomes Z_i take the values 0 and 1. When each constituent decision tree is formed in producing the random forest model, nodes are split based on some impurity metric relating to the outcomes Z_i [53]. For a given node, A, node impurity is defined as

$$I(A) = \sum_{i \in O} g(p_{i,A}), \tag{9}$$

where g is an "impurity" function, O is the set of possible outcomes, e.g. 0 or 1, and $p_{i,A}$ is the proportion of data in A which belong to outcome class i of Z, i.e. 0 or 1. Typical choices for g are the information index ($g(p) = -p \log(p)$) or the Gini index ($g(p) = p(1 - p)$). At a given stage of the decision tree, the splitting variable, and its value, is chosen to maximize the impurity reduction

$$\Delta I_A = p(A)I(A) - p(A_L)I(A_L) - p(A_R)I(A_R) \tag{10}$$

where A_L and A_R are respectively the left and right resulting nodes and $p(A)$ is the probability of A (for future observations) [54]. A variable v's importance in a single tree can be computed as

$$\sum_{i \in P_v} \Delta I_i, \tag{11}$$

where P_v is the set of splits for (i.e. nodes that split on) the variable.

Variable importance (for the entire forest) can similarly be calculated by considering nodes across all decision trees, typically standardized by the number of trees fit or the number of trees where the split was present,

$$VI(v) = \frac{1}{|\text{Number of Trees}|} \sum_{i \in P_v} \Delta I_i. \tag{12}$$

Additional modifications, such as the use of surrogates, can also be added to improve variable importance metrics. The technical details are left to more complete works on random forests, e.g. see [53], and Appendix B.

Variable importance comparison

Although the computed variable importance is a helpful summary statistic for ordering variables in terms of their expected usefulness in predicting the outcome, there are several challenges in using these values to determine which variables are significantly more useful than others. One is that the variables in $\mathbf{C}'$ are d-dimensional proxies of the information derived from the spatial image data. When multiple components are used to describe a single function, i.e. $d > 1$, we must take into account that each individual component in $\mathbf{C}'$ describes only a portion of the associated K function. Therefore, the importance of each component must be combined to describe the importance of each spatial interaction. Yet this importance must be made comparable to that of the meta-variables. Also, we wish to identify spatial interactions and meta-variables that are of "significant importance", which we take to mean that their importance exceeds to a statistically significant degree that of similar variables that are unrelated to the response. This task is complicated by the fact that we are often faced with such a large number of spatial interactions. Given the large number of variables, we expect some to have anomalously large variable importance even when they are independent from the response.

To allow for such comparisons, we standardize the variable importance metrics by adding matched synthetic variables. These synthetic variables are generated by permuting the true data in order to maintain identical distributions, but are independent from the outcome. We denote the synthetic spatial components as $\mathbf{C}'_s$ and the synthetic meta-variables $\mathbf{M}_s$.

A random forest model is fit using both the true and synthetic variables,

$$\widehat{\mathbf{Z}} = f(\mathbf{C}', \mathbf{M}, \mathbf{C}'_s, \mathbf{M}_s) \, . \tag{13}$$

To build $\mathbf{C}'_s$, a random functional variable is selected for each b'th iteration, $1 \leq b \leq B$, where B is a user specified parameter. The d principal components associated with each patient are permuted across patients, resulting in assignment of the d components to a random patient and hence outcome. Note that in doing so, the d components are kept together.

One could use B synthetic variables for each functional variable. However, investigations into the model through extensive simulations has shown that a single group of B synthetics is generally sufficient for the K functions, and additional synthetic K variables do not improve power to identify important variables beyond this.

Although one synthetic group for meta-variables could be used, previous work has shown a tendency for random forests to favor continuous predictors over discrete predictors [55]. The model accounts for this tendency through unique synthetic meta-variable groups. Therefore, $\mathbf{M}_s$ is created by permuting each meta-variable across patients B times.

We use these synthetic variables, $\mathbf{C}'_s$ and $\mathbf{M}_s$, to standardize the variable importance values of the true data and build noise thresholds. In doing so we are able to infer which spatial interactions and meta-variables lead to significant improvements in the model accuracy. The details of this are left to Comparing variable importance of spatial interactions and meta-variables to noise.

Due to the innate randomness in the models, the variable importance values fluctuate between runs and model fits of the random forests. To quantify this, we employ cross-validation (CV).

Let the data be randomly assigned to the F folds such that an indexing function κ: $\{1, \cdots, N\} \mapsto \{1, \cdots, F\}$ indicates the partition to which each p^{th} patient's data are allocated. Denote the fitted model, using the true variables as in the model from Eq (13), with the j^{th} fold removed $\hat{\mathbf{Z}}^{(-j)}(\mathbf{C}', \mathbf{C}'_s, \mathbf{M}, \mathbf{M}_s)$. The estimated CV variable importance for each functional variable c (in both $\mathbf{C}'$ and $\mathbf{C}'_s$), which are described by the previously discussed d dimensional principal components $k_c^{(d)}$, is computed as

$$\mathrm{VI}^{(CV)}(c) = \frac{1}{F} \sum_{j=1}^{F} \sum_{i=1}^{d} \mathrm{VI}_j^{(CV)}(k_c^{(i)}) \, , \tag{14}$$

where $\mathrm{VI}_j^{(CV)}(k_c^{(i)})$ denotes the variable importance estimate from the j^{th} fold-removed model $\widehat{\mathbf{Z}}^{(-j)}(\mathbf{C}', \mathbf{C}'_s, \mathbf{M}, \mathbf{M}_s)$ for component $k_c^{(i)}$, which is the i^{th} component relating to the c^{th} function. For meta-variable m, the variable importance is computed as

$$\mathrm{VI}^{(CV)}(m) = \frac{1}{F}\sum_{j=1}^{F}\mathrm{VI}_j^{(CV)}(m)\,. \tag{15}$$

We quantify the uncertainty in the estimate of the variable importance measure for each variable v, both functional and meta, by calculating its standard deviation across the F folds.

$$\mathrm{SD}(v) = \sqrt{\frac{1}{F-1}\sum_{j=1}^{F}\left(\mathrm{VI}_j^{(CV)}(v) - \mathrm{VI}^{(CV)}(v)\right)^2}\,. \tag{16}$$

This uncertainty estimate is used with variable importance estimates created from the mean of non-cross-validated models. Let $\mathrm{VI}_j(x)$ indicate the variable importance metric from the model in Eq (13), iterate j (where we take $1 \leq j \leq F$ for ease, but each run is on the full data set), then the estimates are computed respectively for functional and meta-variables as

$$\mathrm{VI}(c) = \frac{1}{F}\sum_{j=1}^{F}\sum_{i=1}^{d}\mathrm{VI}_j(k_c^{(i)})\,, \text{ and } \mathrm{VI}(m) = \frac{1}{F}\sum_{j=1}^{F}\mathrm{VI}_j(m)\,. \tag{17}$$

Comparing variable importance of spatial interactions and meta-variables to noise

As mentioned above, due to the use of $d \geq 1$ principal component summaries, we expect that for spatial interactions and meta variables that are independent of the outcome the variable importance of the spatial interactions will typically be larger. As such we use the estimated variable importance values for the synthetic variables to calibrate the variable importance between the spatial interactions and meta-variables. We compute for each of the spatial and meta-variables, respectively c and m, the empirical α quantiles of the variable importance values of the synthetic variables. If $\mathbf{C}^{(s)}$ indicates the set of synthetic functions, i.e. the combined synthetic components, and $\mathbf{M}^{(s,m)}$ indicates the set of synthetic meta-variables matched to meta-variable m, then we set

$$\begin{aligned} Q_C &= \inf\left\{q \,:\, \frac{1}{B}\sum_{c_{\text{syn}}\in\mathbf{C}^{(s)}}\mathbb{1}(\mathrm{VI}(c_{\text{syn}}) \leq q) > \alpha\right\}, \text{ and} \\ Q_{M,m} &= \inf\left\{q \,:\, \frac{1}{B}\sum_{m_{\text{syn}}\in\mathbf{M}^{(s,m)}}\mathbb{1}(\mathrm{VI}(m_{\text{syn}}) \leq q) > \alpha\right\}. \end{aligned} \tag{18}$$

Let $\mathbf{Q}_M = (Q_{M,1}, \ldots, Q_{M,|M|})$ where $|M|$ is the total number of meta-variables. Below we always set $\alpha = 0.95$. Letting $Q_{\text{noise}} = \max\{Q_C, \mathbf{Q}_M\}$, we calibrate the variable importance of each true variable computed from the model in Eq (13) that includes synthetic variables, denoted $v_{\text{true},(13)}$, as

$$\mathrm{VI}_{adj}(v_{\text{true},(13)}) = \frac{Q_{\text{noise}}}{Q_C\mathbb{1}(v_{\text{true},(13)} \in \mathbf{C}) + \sum_{m\in M}Q_{M,m}\mathbb{1}(v_{\text{true},(13)} \in m)}\mathrm{VI}(v_{\text{true},(13)})\,. \tag{19}$$

The cross-validated standard deviations of the variable importance values are similarly adjusted based on the model from Eq (8) as

$$\mathrm{SD}_{adj}(v_{\text{true},(8)}) = \frac{Q_{\text{noise}}}{Q_C \mathbb{1}(v_{\text{true},(8)} \in \mathbf{C}) + \sum_{m\in\mathbf{M}} Q_{M,m}\mathbb{1}(v_{\text{true},(8)} \in m)} \mathrm{SD}(v_{\text{true},(8)})\,. \tag{20}$$

These adjusted variable importance values may then be compared to Q_{noise}. Estimates that go below the threshold Q_{noise} have variable importance values that appear to show no statistically significant relationship with the outcome. Plotting Q_{noise}, the adjusted variable importance values, $\mathrm{VI}_{adj}(v)$, along with their adjusted standard deviations gives a simple way to evaluate at a glance which variables appear to be important in predicting the response. An example plot of this is shown in the right hand panel of Fig 3, where Q_{noise} is plotted as a vertical, red-dotted line, and the variable importance values are plotted (black dots) along with their adjusted standard deviations (surrounding bands).

While Q_{noise} estimates the α percentile of the variable importance values corresponding to outcomes that are unrelated the outcome, the variable importance values themselves can be distorted due to overfitting. In essence as more variables are observed, it is increasingly likely to find a variable that seems to have high importance, despite no true relationship to the response.

To account for this potential effect, we employ an additional variable permutation. In this step, we permute the true variables in addition to the synthetic variables, H times. In each forest, the synthetic variables are again used to align the variables (using the previously computed Q_{noise}), but the remaining, aligned "true" variables do not have any relation to the outcome. In each of these forests based on variables that are independent from the response, $1 \leq h \leq H$, the resulting, adjusted variable importance estimates are ordered,

$$\mathrm{VI}_{adj,h,\ell} = \{\ell^{\text{th}} \text{ largest adjusted variable importance value in forest } h\}\,, \tag{21}$$

ensuring $\mathrm{VI}_{adj,h,1} > \mathrm{VI}_{adj,h,2} > \cdots > \mathrm{VI}_{adj,h,V}$ where V denotes the total number of variables between the spatial interactions, $\mathbf{C}$, and the meta-variables, $\mathbf{M}$. When the number of variables V is large, especially in relation to the number of patients N, we might expect even when the spatial variables and meta-variables are independent of $\mathbf{Z}$ that the largest variable importance values will (far) exceed Q_{noise}. As such, we also compute the α quantile of the variable importance values in each ordered position ℓ for the random forests fit to the permuted data,

$$Q_{\text{int},\ell} = \inf\{q \;:\; \frac{1}{H}\sum_{h=1}^{H} \mathbb{1}(\mathrm{VI}_{adj,h,\ell} \leq q) > \alpha\}\,. \tag{22}$$

We let $\mathbf{Q}_{\text{int}} = (Q_{\text{int},1}, \ldots, Q_{\text{int},V})$, and call it the "interpolation threshold".

We include $\mathbf{Q}_{\text{int}}$ with Q_{noise} in order to quantify overfitting and thereby evaluate the significance of values of the variable importance values $\mathrm{VI}_{adj}(v)$; this is the orange-dotted line in the right hand panel of Fig 3.

In summary, variables with adjusted importance values, $\mathrm{VI}_{adj}(v)$, that are larger than both Q_{noise} and $Q_{\text{int},\ell}$ exhibit importance that significantly exceeds (at the $1 - \alpha$ level) what we might expect from similar variables that are unrelated to the outcome. This holds taking into account the inflation in the variable importance values that arise from fitting the random forest to a large number of spatial interactions and meta-variables.

Choice of parameters B and H

The parameters B and H must be selected by the user. In constructing Q_{noise} and $Q_{\text{int},\ell}$, these values define the number of synthetic variables used to approximate the quantiles in Eqs (18) and (22), respectively. We have explored several choices of B and H in our simulation experiments, and found that when the level of the thresholds is set at $\alpha = 0.95$, the values $B = H = 100$ give satisfactory results for the up to 16 cell types considered. Further details of these simulations are shown in Appendix C. In summary, we have observed that for up to 16 cell phenotypes and with $\alpha = 0.95$, values of B and H exceeding 50 behaved almost identically in terms of their false positive and false negative rates for identifying important variables. There was no apparent advantage, or apparent disadvantage other than additional computational burden, for choosing larger values of B and H for the number of cell interactions considered. We note that we only considered up to 16 cell phenotypes. We recommend using larger values of B and H for a larger number of cell phenotypes.

Predictive accuracy estimates

In weighing the significance of the computed variable importance values, one should also consider the overall predictive accuracy of the final model for the outcomes. For example, a variable may have a large variable importance value within a model that does not lead to improved predictions of the outcome over naïve models.

We consider out-of-bag accuracy (OOB) to estimates how well our final random model works on unseen data, and compare it to the naïve approaches that we label GUESS, and BIAS.

OOB is computed by predicting the data left out during each CV iteration,

$$\mathrm{OOB} = \frac{1}{N}\sum_{p=1}^{N} \mathrm{Diff}\left(\hat{Z}^{(-\kappa(p))}\left(\mathbf{X}^{(-\kappa(p))}, \mathbf{M}^{(-\kappa(p))}\right), z_p\right), \tag{23}$$

for the N patients and where Diff indicates a difference function. For classification problems such as with the TNBC or LUSC versus LUAD data, this may be defined using an indicator function, $\mathrm{Diff}(x, y) = \mathbb{1}(x = y)$.

GUESS is defined as the probability of correctly guessing a patients outcome by randomly selecting the outcome based on the outcome's observed frequency in the original data. If for the observed frequencies of outcomes $1, \ldots, n$ in the data are $p_1, \ldots, p_n$, this is computed as

$$\mathrm{GUESS} = \sum_{i=1}^{n} p_i^2 \tag{24}$$

For example, in the TNBC data in which we wish to predict the "compartmentalised" versus "mixed" result with a proportion p of "compartmentalised" patients, this amounts to computing the rate at which we would accurately guess the outcome by flipping a coin independently for each patient with probability p of heads, and guessing the outcome is "compartmentalised" for heads, and is "mixed" otherwise. The GUESS value would therefore have a success probability of $p^2 + (1 - p)^2$.

BIAS is built by always guessing the patients outcome that is the most prevalent outcome in the sample. For classification problems, the most likely outcome can naturally be defined by the data mode. We compute

$$\mathrm{BIAS} = \frac{1}{N}\sum_{p=1}^{N}\mathrm{Diff}\Big(\mathrm{mode}(\mathbf{Z}) - z_p)\Big). \tag{25}$$

For two outcome data sets like the TNBC or lung cancer examples, this will have success probability for a random patient drawn from the sample of max{p, 1 − p}.

When no signal is in the data, we expect OOB to perform similarly to GUESS or BIAS. However, if the covariate information available is useful in predicting the response, we expect OOB to far exceed the naïve estimates. The variability of OOB estimates is further analyzed and discussed in Appendix D.

Variable importance plot

The variable importance plot summarizes the variable importance values of both the spatial interactions and meta-variables. It shows how they compare to what we might expect from similar noise variables which are unrelated to the outcomes, and also summarizes the overall efficacy in predicting the outcomes using the random forest model. Fig 4 is the variable importance plot created from simulated data with two cell types, *A* and *B*, and two meta-variables with differing distributions, age and sex. Variables are displayed according to their adjusted variable importance estimates (black dots), standardised with respect to the variable with the

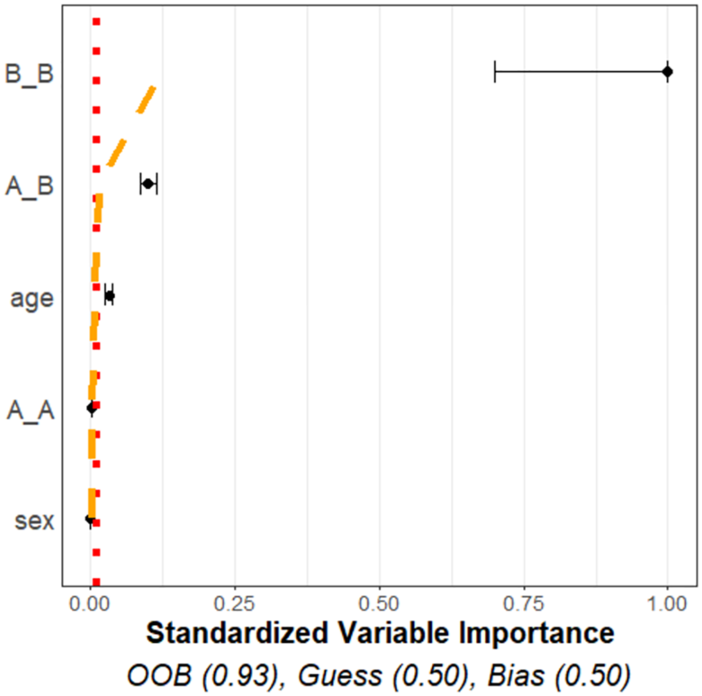


Fig 4. Sample variance importance plot. This sample variable importance plot uses simulated data with a binary outcome, two cell types, and two meta-variables. The data was simulated with significant differences between the outcomes in the *B_B*, *A_B* spatial interactions, and age meta-variable, but no significant difference across sex and the *A_A* spatial interaction. The point estimates of the variable importance values are the black dots, with accompanying intervals indicating the uncertainty. The red dotted straight line is the noise threshold and the orange dashed curved line is the interpolation threshold. Both thresholds are used to indicate if a variable is predictive of the outcome beyond that of random noise. The variable importance values of the known significant variables are shown to exceed that of the noise and interpolation thresholds.

https://doi.org/10.1371/journal.pcbi.1011361.g004

largest variable importance (such that the standardised variable importance values range from 0 to 1). The variables are listed on the y-axis in order of decreasing variable importance with the largest values being at the top.

The plot visualizes uncertainty in computing the VIs through (gray) intervals representing one standard deviation on either side of the estimate. These intervals are the adjusted standard deviations estimated through CV, as given in Eq (20).

The noise threshold Q_{noise} as well as the interpolation threshold $\mathbf{Q}_{\text{int}}$ at 95% levels ($\alpha = 0.05$) are also shown. Fig 4 presents the noise threshold as a red, dotted, vertical line and the interpolation threshold as a orange, dashed, curved line.

Estimates for predictive accuracy are given at the bottom of the figure. The estimates include OOB, GUESS, and BIAS, respectively defined in Eqs (23), (24) and (25).

Further interpretation of the variable importance plots are given in the sections Simulation study and Applications to MIBIscope data. These sections consider simulations for which there is a known solution and real data examples that can be interpreted.

Results

In this section we present results on simulation experiments and analyses of two real experimental data sets using the proposed methodology. All investigations were completed on a standard personal laptop (with an Intel i7 processor), and did not require high-end computational resources. In fact, the computational complexity can be readily shown to be

$$O(N\log(N)[dc+m][F+H]BT)\,, \tag{26}$$

and notably increases only linearly in each variable aside from the sample size, which includes an additional logarithmic factor. Additional details on the computational complexity of random forests models can be found in e.g. [53].

Many additional (unreported) simulations were also conducted to test the robustness of the method against various patterns, image shapes, sample sizes, and parameter choices. We found robustness to sample size and variable count (See Appendix E), informative and non-informative patterns, image shape and size (provided images are large enough to capture the desired relationships), and parameter choices (See Appendix C).

Simulation study

We present the results here of simulation experiments in which we applied the proposed methods to simulated spatial point patterns. In particular, we produced simulated point patterns with properties, such as cell counts, numbers of phenotypes, etc., similar to that of the TNBC data set in [12]. The real data motivating this simulation experiment are analysed in the following section. The primary goals of the simulation experiment were to demonstrate that the proposed method is effective at controlling the false discovery rate, i.e. the rate at which we mistakenly identify spatial interactions or meta-variables with no relationship to the response of interest as being informative or important, and also that it has the power to detect important spatial interactions.

We considered simulated data from 34 "patients". These patients were defined as being negative or positive for a binary outcome Z (note that positive/negative here refers to the arbitrary outcome Z, and is not related to hormone receptor status as it does in the term TNBC). We let there be 17 positive and 17 negative patients and simulate one image per patient. Each image consists of a point pattern with 16 cell phenotypes. Patients were also ascribed a single meta-variable, which we call age. We developed a random forest model as described in Materials and methods to predict patient outcome using interactions between the 16 different cell types in the images and the additional meta-variable.

We considered two main settings: (1) a simulation with only non-informative variable-outcome relations and (2) a simulation with both informative and non-informative variable-outcome relations.

To generate the images, the cells were placed according to multiple, potentially nested, (modified) Thomas processes, which are constructed iteratively [56]. For a given image, a Thomas process first places cells (of a given type, say $c1$) at random, according to a Poisson process. As such the number of cells to place, n_{c1}, is determined at random based on a Poisson distribution, where the distributional parameter is user selected. For our simulation study, these are selected to correspond to the mean number of cells in the TNBC dataset. We standardize the images to unit length in both the x and y directions.

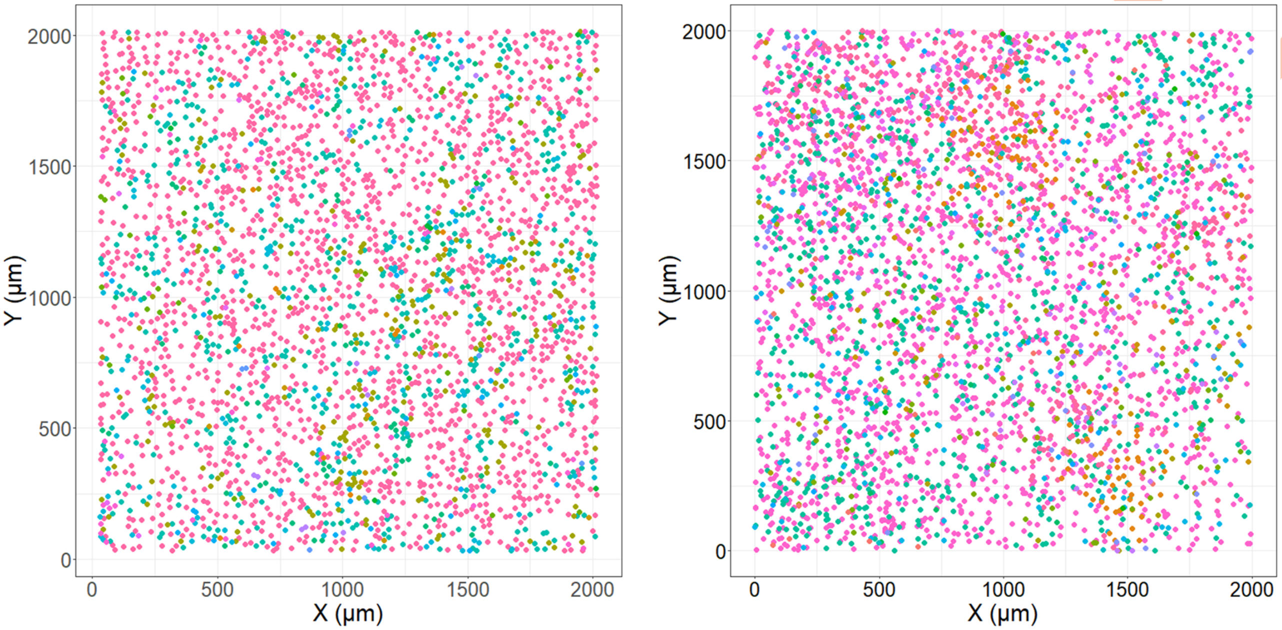
Around each cell $c_{a,c1}$, $a \in \{1...n_{c1}\}$, cells of a different type, say $c2$ may be placed. Again the number of $c2$ cells are randomly selected based on a Poisson distribution with a user selected parameter. The coordinates of the $c2$ cells are placed according to another distribution in such a way that they either cluster or disperse around the $c1$ cells. In our experiments, a bivariate Normal distribution is used to place the $c2$ points, so that the mean coincides with the location of a randomly selected $c1$ cell, and the covariance matrix is a scalar multiple σ^2 times the identity matrix. By varying σ^2 in the outcome groups, clustering or repulsion in the cell interactions can be introduced.

Additional cells can be simulated around the points of $c1$ or $c2$, and so on. Moreover, the original cells can be removed such that the new cells appear to exhibit self-clustering. Compilations of such placement patterns, with use of potentially different distributions, can achieve images with varying degrees of clustering or regularity.

In our simulations, 16 cells were iteratively placed according to this modified Thomas pattern. Some cells were placed completely at random ($c1$, $c4$, $c5$, $c6$, $c7$, $c13$, $c14$, $c15$), some were placed exhibiting self-clustering ($c8$, $c9$, $c10$, $c11$, $c12$), and some were placed exhibiting clustering around $c1$ ($c2$, $c3$, $c16$). In what we call the "no–relation" simulation, the cells locations were simulated in the same way for both the positive and negative patients. However in the "relation" simulation, $c2$ exhibited increased clustering around $c1$ while $c3$ exhibited repulsion from $c1$, for positive outcomes. Similar to the true data, some cell types were present hundreds of times per image while others only rarely appeared. Each of the synthetic cell types were generated to mimic behaviors and frequencies seen in the TNBC data. Fig 5 presents two images, an image from the TNBC data set and a simulated image.

In this way a single image for each of the 17 positive and 17 negative patients were simulated. The meta-variable age was simulated as a Normal random variable with unit variance. While in the no-relation case the mean age was constant, 25 for both outcomes, in the relation-case the mean age was set to be 25 for negative outcome patients and 27 for positive outcome patients. Moreover, in the relation-case, 2 of the 17 positive patients were given no-relation to the outcome as additional "noisy" images.

When modeling both cases, there were several tuning parameters selected. We used the standard choices of 500 trees for each random forest, each tree using 80% variable selection and full data bootstrap, 10 folds in cross-validation, and a standard significance level of α = 0.05. We used 100 interactions for the permuted random forest in creation of the interpolation cut-off. Appendix C highlights several numerical investigations of these values, showing robustness in the results for these choices. Creation of the K functions also required selection of the maximum radius, R. Although we investigated the effect of the choice, we saw little variation in the results and used a traditionally recommended 25% of the side length of the image, along with the previously discussed isotropic edge corrections for the simulations. Moreover,



(a) TNBC data.

(b) simulated data.

Fig 5. Comparison of TNBC and simulated data. (a) An image from the TNBC data and (b) an image from the simulated data. Different colors indicate one of the 16 different cell phenotypes, showing the comparability of the simulations and true data.

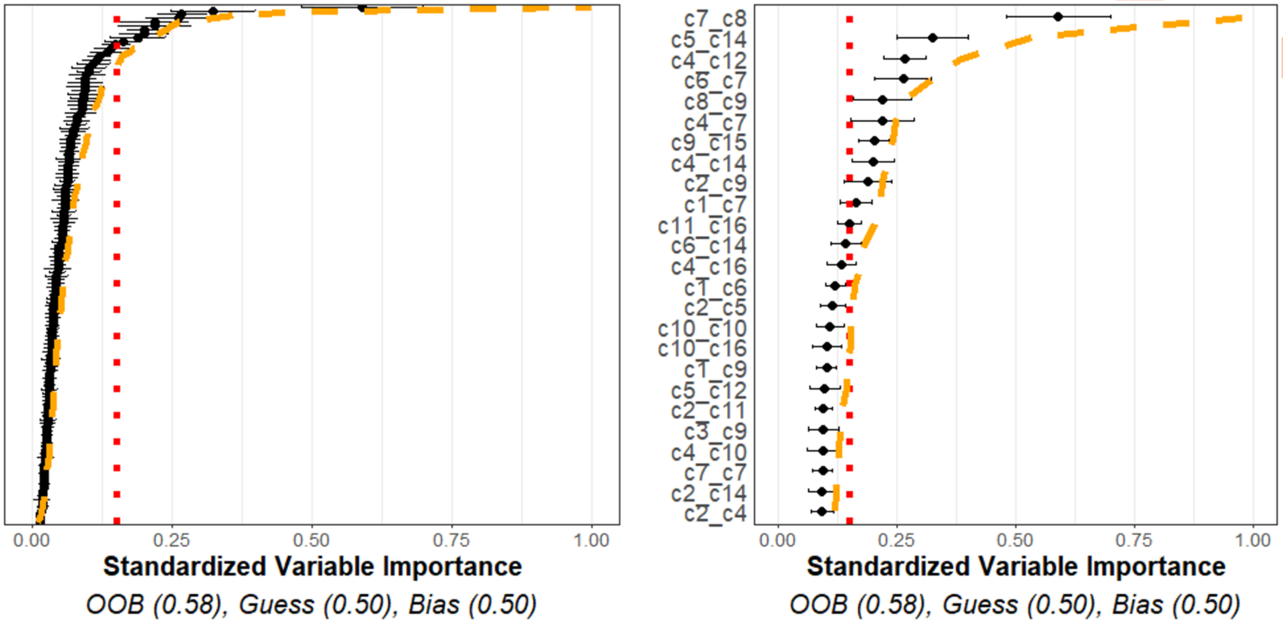
https://doi.org/10.1371/journal.pcbi.1011361.g005

the *K* functions were summarized using 3 principal components, which were selected to match the TNBC analysis. In the TNBC case 3 components explain at least 95% of the total variance explanation for each *K* function.

Fig 6a shows representative variable importance for all variables and Fig 6b shows the top 25 in the "no–relation" case. One may see from the plot that some variable importance estimates exceeded the red noise threshold, but they all were observed to be below the orange interpolation threshold. This may be interpreted that the observed variable importance values did not exceed, to a significant degree, what we would expect to see from similar variables that are known to be independent of the response. Moreover, the prediction accuracy estimate (OOB) indicates the model performs similarly to a naïve guessing approach. Taken together, the plot indicates that none of the spatial interactions appeared to be important in predicting the outcome, as was expected in this case.

On the other hand, Fig 7 shows a representative variable importance plot computed from a single simulation run in the "relation" simulation. All variables are shown in Fig 7a, and only top 25 are shown in Fig 7b. Although many variables are still below the noise and interpolation thresholds, the known related variables are found to have a significant relationship with the outcome variable. Moreover, the OOB estimate far exceeds that of the GUESS or BIAS estimates. This plot indicates that interactions between the $c1$, $c2$, $c3$ cell types, and the age variable, appeared to be useful, to a statistically significant degree, in predicting the outcome, once again as expected. We note that since $c2$ and $c3$ cells are distributed around $c1$ cells, any change in these distributions will necessarily lead to differences in the cross *K* functions between these two cell types as well, as observed.

In order to verify that each of the orange and red lines appeared to be appropriately calibrated, we performed an additional simulation experiment. We considered two cases (1) with



(a) Full variable importance plot.

(b) Top 25 variable importance plot.

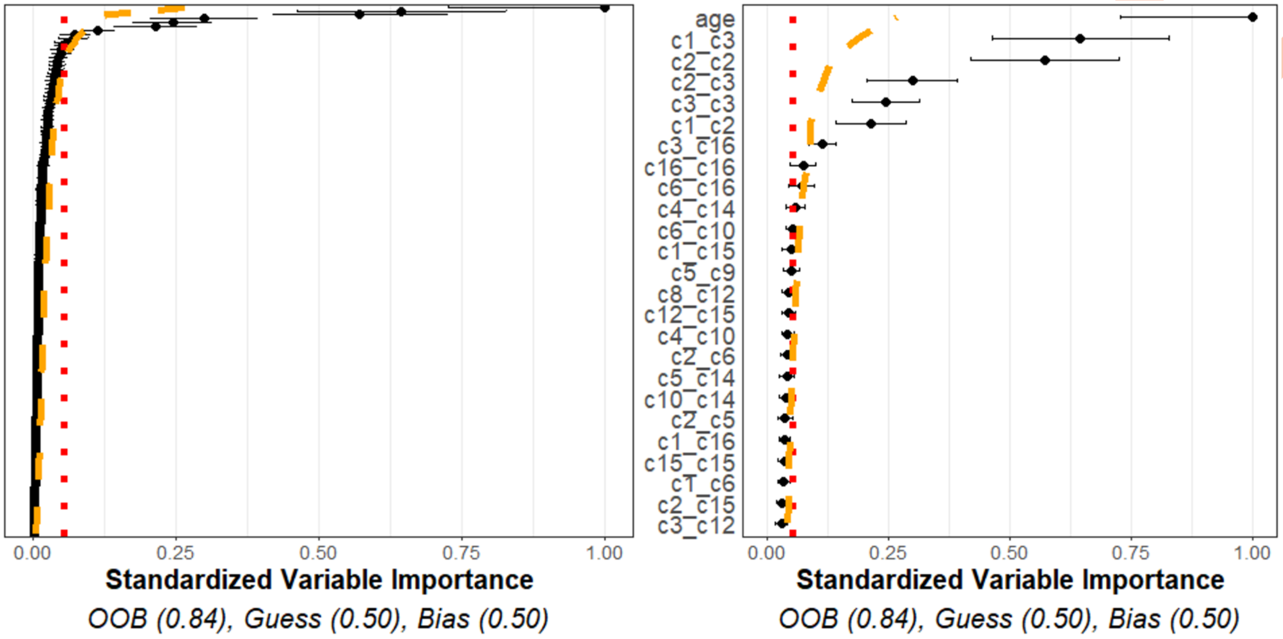
Fig 6. No relationship simulation. Simulation of 16 cell types for 34 patients with meta-variable age. (a) Figure with the variable importance values for all variables. (b) Figure with only the top 25 largest variable importance values. All variables were generated with no-relationship to the outcome and all were determined to have no relation to the outcome beyond noise as the variable importance estimates are below noise and interpolation thresholds.

https://doi.org/10.1371/journal.pcbi.1011361.g006

4 cell types and (2) with 16 cell types. Each cell type other than cells *c*1 and *c*2 are generated such that they have no relationship to outcome. We modify the clustering of *c*2 around *c*1 in the positive group, and examine the rate at which the variable importance estimated for the *c*1_*c*2 spatial interaction exceeded the various thresholds (red/orange lines). By changing the standard deviation in the Thomas process for placing *c*2 points around *c*1 points, we were able to investigate whether the approach is able to detect the presence of a relationship when the cells either cluster or are more dispersed across the binary outcomes. The resulting power curves, based on 100 simulations for each setting, are shown in Fig 8. These show the rate at which the variable importance estimates exceed the 95% noise threshold, interpolation threshold, and both the noise and interpolation thresholds. In reference to Fig 8, the underlying standard deviation in the Thomas process relating c1 and c2 takes the value of 0.025 in the "no relation" case. As such, for this value it is desired that the noise and interpolation thresholds are exceeded no more than $\alpha = 0.05$ proportion of the time. Otherwise, detecting that standard deviation values smaller/larger than 0.025 lead to increased/reduced clustering of *c*1_*c*2 in the positive groups is desired.

We saw that each threshold appeared to be effective in detecting clustering or dispersion relationships, such that even relatively small changes were detected with high frequency, even when considering many variables, see also Appendix E.

Table 1 further shows the empirical false discovery rate computed from 100 independent simulations for 4 and 16 cell types with a simulated population of 34 patients evenly split between positive and negative response groups. In this case all spatial interactions and meta variables were simulated independently of the response. The false discovery rate is computed as the percentage of simulations for which **any** VI exceeded the corresponding threshold or



(a) Full variable importance plot.

(b) Top 25 variable importance plot.

Fig 7. Relationship simulation. Simulation of 16 cell types for 34 patients with meta-variable age. (a) Figure with the variable importance values for all variables. (b) Figure with only the top 25 largest variable importance values. Most cell types were generated with no relationship to the outcome. However, age, *c1_c2*, and *c1_c3* were designed to have a relationship with the outcome (which naturally means *c2_c2* and *c2_c3* would also have relationships to the outcomes). These variables are seen with significantly larger variable importance values than the thresholds and other variable importance values.

https://doi.org/10.1371/journal.pcbi.1011361.g007

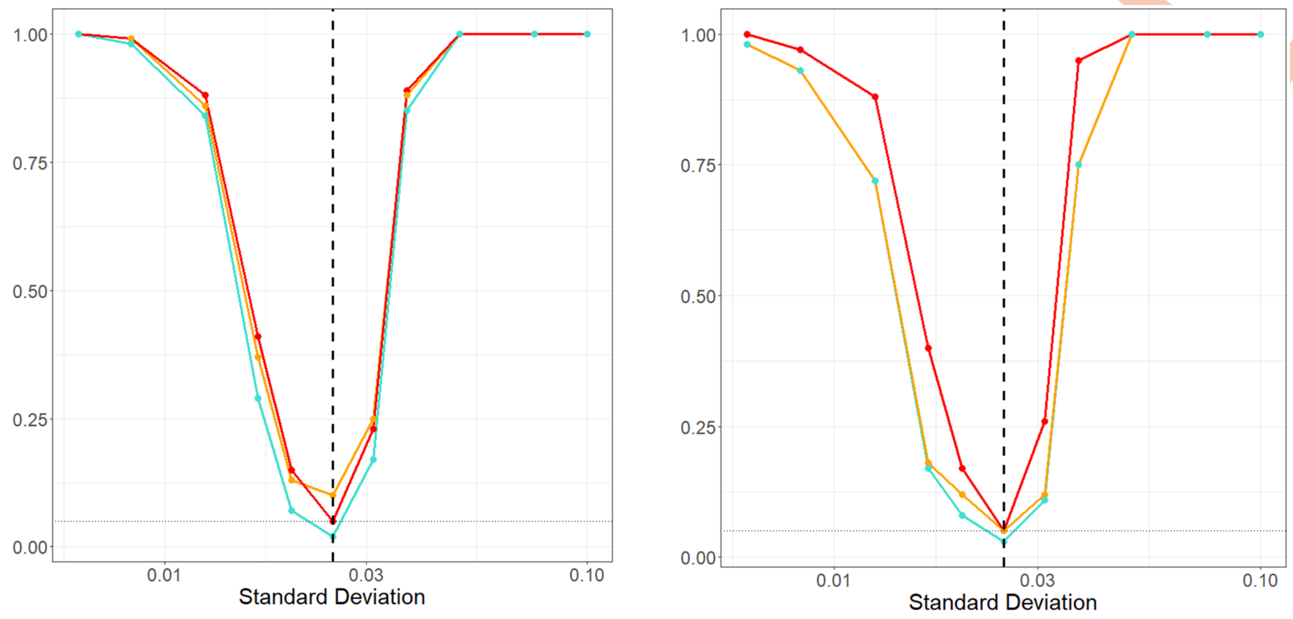
both thresholds. We also considered “Variable importance—1 SDCV”, in which the false discovery rate is computed as the percentage of simulations in which any of the cross-validated one-standard-error intervals for the VI lies to the right of the corresponding threshold, and “Largest Variable Importance”, in which we only considered whether the largest VI computed exceeded the corresponding threshold.

These results suggest that while either of the two thresholds alone are not suitable to control the false discovery rate, comparing the VI along with the one standard error cross-validation interval to *both* thresholds (red and orange lines) yielded a false discovery rate very close to the nominal rate. Moreover, comparing the largest VI to both thresholds controlled the rate of falsely identifying the spatial interaction or meta-variable associated with that as being statistically important.

In Appendix E, we provide additional simulation evidence illuminating the effect of increasing the number of cells considered. These results suggested that the the false discovery rate is not influenced by the number of cells considered, and remains controlled. The power to detect a single, important interaction was observed to decrease as the number of cells considered increased, as expected.

Applications to MIBIscope data

In this section we present two applications of the proposed methods to MIBIscope data sets. The first investigates known clusters in tumors related to triple negative breast cancer tumors, while the second investigates unknown relations in tumors related to lung cancer. Unless



(a) 4 cell types.

(b) 16 cell types.

Fig 8. Power curves. Power curves showing the empirical rate, from 100 simulations, that the variable importance for the spatial interaction *c*1_*c*2 exceeded the 95% noise threshold, interpolation threshold and both the noise and interpolation thresholds, indicating a significant spatial interaction is detected. The curves indicate the threshold method: above both thresholds (teal), above the curved interpolation threshold (orange), and straight noise threshold (red) (the colours correspond with the noise curves used in the variable importance plots). The *x* axis gives the standard deviation parameter controlling the distribution of *c*2 cells around *c*1 cells. The vertical black dotted line is the base case, when both classes exhibit the same interactions across all cell types, including the *c*1_*c*2 interaction. To the right of this line, *c*2 cells are less densely clustered around *c*1 cells, and to the left of the line there is increased clustering around *c*1 cells. The light horizontal, dotted gray line indicates the desired mis-classification rate when no signal is present (i.e. 0.05). (a) Spatial data with 4 cell types is used to create the curves. (b) Spatial data with 16 cell types is used to create the curves. Both images show the method is effective at correctly detecting when the important interaction does and does not differ between the patient outcomes. This is seen as all lines quickly climb to 1 (perfect detection of a signal) as the standard deviation parameter moves further from the no effect case (vertical line). The size and power is similar between the simulations despite the large increase in total interactions considered.

https://doi.org/10.1371/journal.pcbi.1011361.g008

Table 1. Table showing empirical false discovery rate based on 100 independent simulations for both *T* = 4 and *T* = 16 cell types with 34 patients split into positive and negative groups. The false discovery rate is computed as the percentage of simulations for which any VI exceeded the corresponding threshold or both thresholds (Variable Importance Only), any of the cross-validated one standard error intervals for the VI lies to the right of the corresponding threshold (Variable Importance-1 CVSD), and whether the largest VI computed exceeded the corresponding threshold (Largest Variable Importance).

	Empirical False Discovery Rate:		
	Nominal Rate 5%		
Threshold:	**Noise**	**Interpolation**	**Both**
Variable Importance Only			
T = 4	0.35	0.15	0.07
T = 16	1.00	0.74	0.09
Variable Importance- 1 CVSD			
T = 4	0.23	0.06	0.04
T = 16	0.99	0.05	0.04
Largest Variable Importance			
T = 4	0.35	0.05	0.05
T = 16	1.00	0.03	0.03

https://doi.org/10.1371/journal.pcbi.1011361.t001

otherwise stated, all parameter values for the model discussed previously remain the same, see Simulation study.

Triple negative breast cancer. We investigate the TNBC data set in [12]. This data was obtained via a MIBIscope, and the authors employ a mixing score to categorise tumors based on their TME. Although we focus on their mixing score, additional information is also available in [12]. The mixing score they use was defined as the proportion of immune cells touching tumor cells, and was calculated as the number of immune-tumor interactions divided by the number of immune-immune interactions for an image. They separate tumors into "compartmentalised" and "mixed" groups (and a "cold" group that we ignore), such that compartmentalised tumors tend to have tumor cells aggregated together with immune cells located around or away from the tumor cells, and mixed tumors tend to have tumor cells and immune cells mixed together. This makes a useful test data set on which to employ our method, as it provides two tumor groups that explicitly have different TMEs. The data set contains 33 patients with a single image per patient, 18 mixed tumors and 15 compartmentalised tumors. We define the outcome Z as 0 for mixed and 1 for compartmentalised, and we wish to predict the outcome using interactions between 16 cell phenotypes in the images and an additional meta-variable, age.

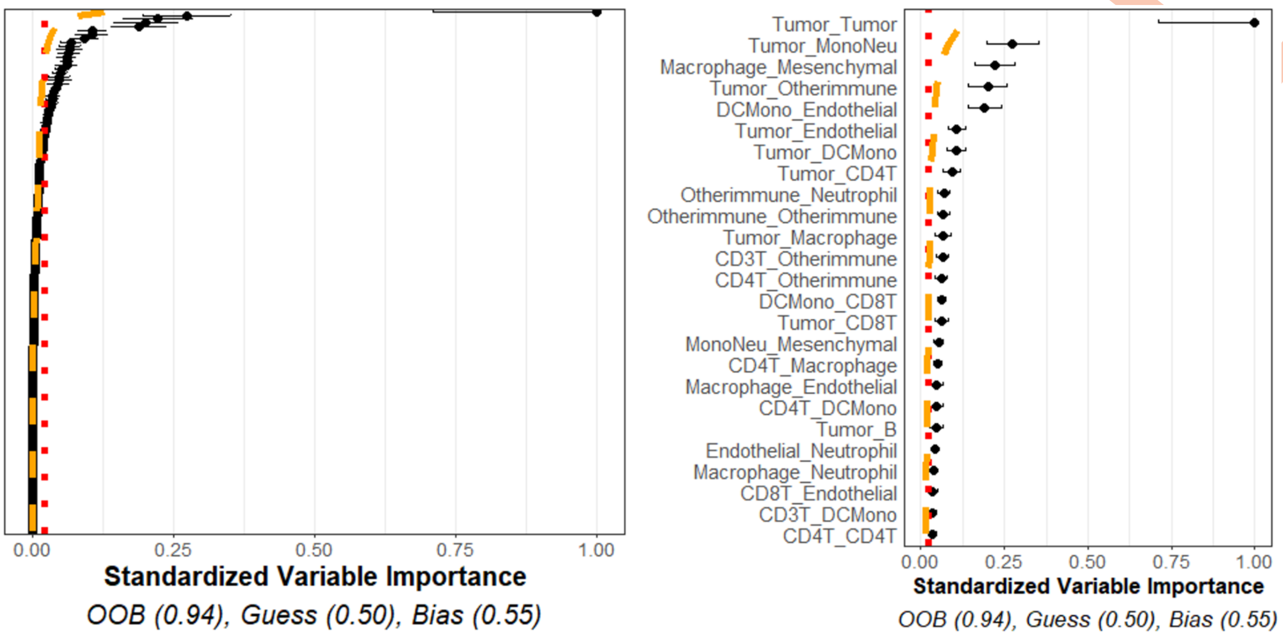
One tuning parameter with K functions is the maximum radius R. For this data we investigated using several options–25, 50, 100, 500, and 1000 micrometers. In all cases the conclusions were comparable, illustrating a robustness to the choice. The only (expected) difference is that as R increases, additional principal components, d, are recommended to capture the variations in the functions. With this observation and domain knowledge, K functions up to a maximum radius of 50 micrometers were used.

The results of this analysis are shown in Fig 9. As expected, many interactions are shown to be non-significant. However, there are a reasonable number above both thresholds and the predictive accuracy estimates agree that there are some important relationships in the data. Specifically, tumor cell-tumor cell interactions came as the top interaction, consistent with the characteristics of the "compartmentalised" tumors where few immune cells infiltrate the tumors and tumor cells are densely packed.

We also applied the method introduced in [31], which is comparable to fitting a functional generalized linear model (logistic regression) to the data and evaluating for the parameter significance corresponding to each interaction. When all interactions are included in such a model, no significant variables at 5% level were identified. After sequentially removing the least important variables based on the variable importance ranking supplied by `funkycells`, this approach was only able to detect the significance of the Tumor_Tumor interaction when it was the only variable included in the model.

Although the method we present here can indicate important relationships, it does not quantify the type of differences. While meta-variables can be easily compared using traditional statistical methods, the K functions are more difficult to analyze. To this end, we can also consider plots of the K functions. Fig 10 examines the Tumor_Tumor interaction, which is found to be significant, and the CD4T_Endothelial interaction, which was found to be insignificant. The significant K functions seem to be well separable (i.e. K functions are clustered with K functions having the same outcome), while the insignificant K functions are not easily separable between outcome groups, and have high variability in individual K functions of the same group.

Lung adenocarcinoma versus lung squamous cell carcinoma. We also applied our approach to attempt to predict different pathological subtypes of non-small cell lung cancer. In this section, we aim to measure to what degree the TME of two of the most common subtypes of lung cancer, LUAD and LUSC, can be differentiated using the spatial relationships



(a) Full variable importance plot.

(b) Top 25 variable importance plot.

Fig 9. TNBC variable importance. Variable importance plot and random forest model summary for predicting "compartmentalized" versus "mixed" tumor types with the TNBC data. (a) Figure with the variable importance values for all variables. (b) Figure with only the top 25 largest variable importance values. The OOB far exceeds those of naïve models, and many of the spatial interactions between tumor cells and immune cell populations exhibited significant variable importance values, suggesting important interactions in the data (such as Tumor_Tumor).

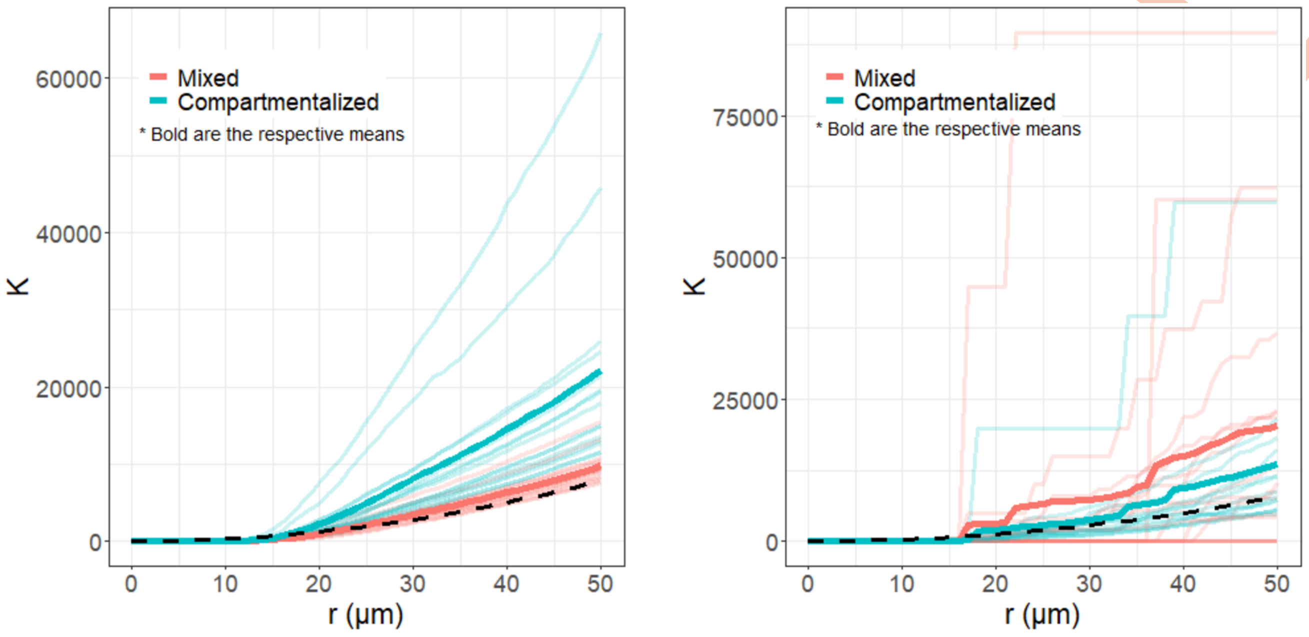
https://doi.org/10.1371/journal.pcbi.1011361.g009

between phenotyped cells as characterised by *K* functions. Our dataset contained 44 LUAD and 20 LUSC samples stained with antibodies against 35 proteins to enable the phenotyping of tumor cells, fibroblasts, and 10 immune cell subsets after scanning on the MIBIScope. The cells were pre-processed and classified and a summary of this data is given in S1 Table. *K* functions were computed between each cell type with again a maximum radius of 50 micrometers.

The variable importance plot and model summary when using our method to predict LUAD versus LUSC are shown in Fig 11. The results suggest that no spatial interactions are significantly useful in distinguishing LUAD versus LUSC. The OOB accuracy was observed to be on-par with a naïve method, and none of the variable importance measurements exceeded the 95% quantile of what we would expect from independent point patterns. This indicates our method using homogeneous *K* functions applied to these specific cell phenotypes is unable to differentiate between the TME of LUAD and LUSC cancer types using this data.

For the purpose of comparison, we also fit functional linear models akin to those proposed in [31] to this data. When all interactions were included, none of the linear model parameters were significant at the 5% level. After sequentially removing the least important variables down to the top 10 based on the variable importance ranking supplied by `funkycells`, we interestingly observed a, apparently spurious, significant interaction associated to spatial distribution of CD8Tcells. In an apparent contradiction, such models based on reducing the number interactions further, including examining a model with only the CD8Tcells_CD8Tcells interaction, all suggested no significant interactions.

This analysis is presented to demonstrate the application of the above methods to real data, and emphasize that while naïve or repeated applications of existing methods might lead to



(a) Significant K function.

(b) Non-significant K function.

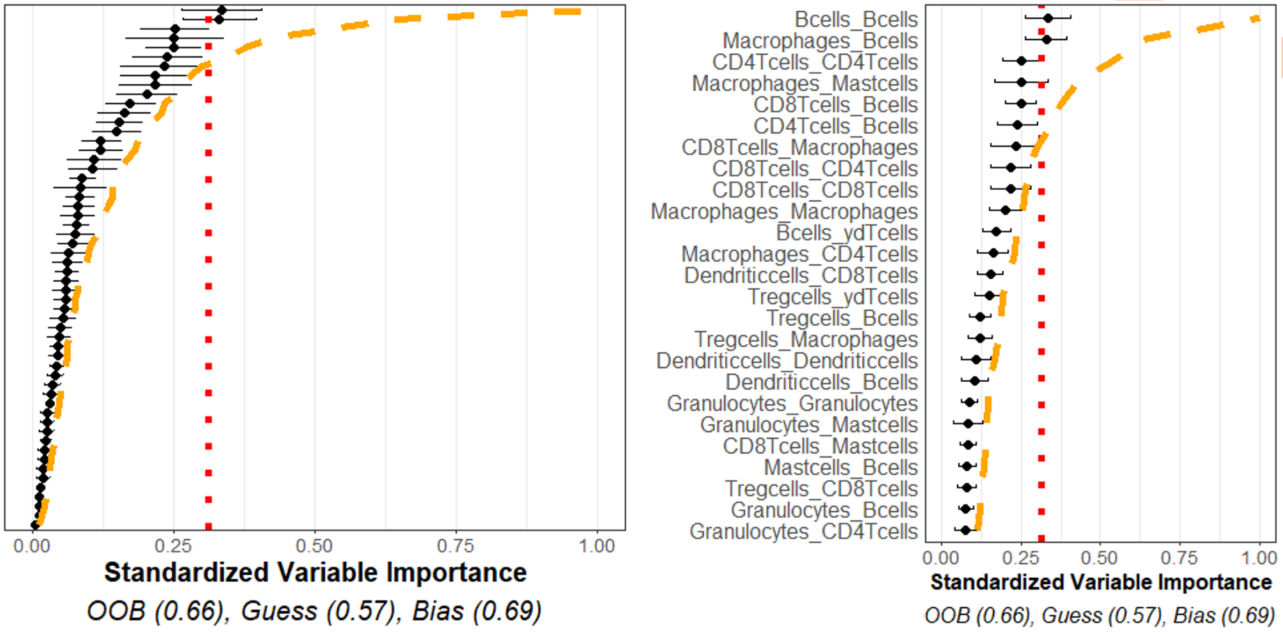
Fig 10. Example TNBC *K* functions. The *K* functions from the different outcomes are compared in these two plots. In both plots, the *x*-axis indicates the radial distance, *r*, in micrometers and the *y*-axis is the value of the *K* function. The lightly colored lines are the *K* functions for each patient, while the bold lines indicates the average (point-wise mean). In the figures, red indicates the mixed tumors while blue indicates the compartmentalized tumors. The black dashed line indicates the curve of a totally spatially random process for reference. (a) Plots the *K* functions for the Tumor_Tumor interaction, which was found to have significant differences in the outcomes. (b) Plots the *K* functions for the CD4T_Endothelial interaction, which was found have no significant differences between the outcomes. In (a), as expected, the compartmentalized group has relatively larger *K* functions–indicating increased clustering–and the functions are well grouped together. Conversely, (b) shows no clear differences between the *K* functions of the two groups and *K* functions are generally surrounded by *K* functions from patients of an assortment of the groups. That is, *K* function patterns vary widely even within the same outcome groups.

https://doi.org/10.1371/journal.pcbi.1011361.g010

spurious findings, our approach is built to control the rate of false discovery of important spatial interactions in the TME. The lack of significance in this example suggests though the need to consider a variety of new methodological approaches when analysing complex TME data. It also highlights the potential limitation of modeling cell interactions through K-functions, which intrinsically assume a degree of spatial homogeneity in the spatial distribution of cells. These issues are discussed further in the following section.

Discussion

In this paper, we present a novel method for the analysis of data-rich TME spatial data. We consider the case in which we compare TME and meta data or clinical data from two different patient groups, and develop a model to identify significant differences between these groups in the distribution of cell phenotypes, protein antigens, or a mixture thereof. In addition, the model aims to predict which group a new patient is in using their TME and meta data. Our model employs a combination of spatial statistics and functional data analysis and is applicable to marked point processes in general. Benefits of the model include general applicability, with few tuning parameters, and easily visualised and interpreted output. We find our model to be robust to the choice of the tuning parameters. The model demonstrates a powerful ability to



(a) Full variable importance plot.

(b) Top 25 variable importance plot.

Fig 11. Lung cancer variable importances. Variable importance plot and random forest model summary for predicting LUAD versus LUSC tumor types. (a) Figure with the variable importance values for all variables. (b) Figure with only the top 25 largest variable importance values. The OOB is similar to a naïve model, and none of the measured variable importance value were statistically significant, indicating no significant variable interactions or meta-variables.

https://doi.org/10.1371/journal.pcbi.1011361.g011

identify important variables while maintaining good predictive power. It is also resistant to overfitting and manages false discovery rate.

We evaluate our approach on simulated data, demonstrating the effectiveness of our method for marked spatial point patterns with known interactions. Our approach is then applied to TNBC and lung cancer data obtained from multiplexed ion beam imaging. For the TNBC data, we compare two groups of tumors that are separated into "compartmentalised" and "mixed" groups based on the degree of tumor-immune cell interactions. This separation of tumors gives two groups with explicit differences in their TMEs, making a useful data set for the demonstration of our model. The model demonstrates good predictive accuracy, and identifies expected specific cell phenotypes relationships, and interactions of interest. The lung cancer data shows our model can also detect lack of differences in cell phenotypes relationships.

Throughout this paper we have assumed homogeneity of the underlying spatial point processes in defining the K functions used in the model. Spatial homogeneity is defined such that the intensity of a given mark is independent of spatial location [14–16]. Whilst such an assumption may be reasonable in some cases, given the complexity of the TME, homogeneity may not always apply. We note, however, that since our approach compares K functions between different groups, rather than against the theoretical K functions associated with complete spatial randomness (as is typical in other circumstances), the lack of underlying spatial homogeneity in the TME for a data set may not be overly problematic. Statistical tests for homogeneity exist [14–16]. Inhomogenous K functions can be used in an attempt to mitigate the issue of inhomogeneity [14–16, 57]. Regardless of how K functions are defined they are amenable to being used in our methodology. Furthermore, whilst we have focused throughout this paper on *K* functions, we note that the methods presented here can be applied to any summary functions from spatial statistics, or indeed any functions in general. See, for example, suggestions in [24, 58, 59].

In addition, we note that $\mathbf{a}_{c,t}^{(p,i)}$ in Eq (1), may consist of the raw measured protein expression level for each protein, and may also include other cell information (e.g. cell size). Our approach could potentially be adapted for analyzing raw protein expression levels via the use of mark-weighted *K* functions [14, 60]. Consequently, the approach may be useful in methods applied to cell phenotypes based on proteins (OPAL, MIBI, Phenocycler Fusion) or transcripts (Xenium, Cosmx, MERscope) expression [43–49]. Usage in non-cellular contexts is also possible.

We have also assumed in this paper outcomes which are categorical, or classes in a group. In this way, we say the model performed classification. Although we considered two classes, a larger number of classes is directly possible. Further extension of this method to real-valued data, e.g. survival time, is likewise natural. Random forest models designed for continuous or survival time analysis exist, e.g. [52, 61, 62], and metrics such as L_2 error can be used in place of the difference function of the OOB and naïve estimates.

Supporting information

S1 Text. Appendices. Appendices A-E, detailing supporting details regarding edge corrections (A), surrogate splitting methods (B), as well as empirical evidence from simulations supporting choice of *B* and *H* parameters (C), investigations of summary metrics OOB, BIAS, and GUESS (D), and consideration of performance under low sample, high interaction count setting (E).
(PDF)

S1 Table. Lung Cancer Summary Table. Table summarizing the patients, number of images, type of lung cancer, and the cell count for each type.
(PDF)

Author Contributions

Conceptualization: Marie-Liesse Asselin-Labat, Gregory Rice, Jack D. Hywood.

Data curation: Claire Marceaux, Kenta Yokote, Marie-Liesse Asselin-Labat.

Investigation: Claire Marceaux, Kenta Yokote, Marie-Liesse Asselin-Labat.

Methodology: Jeremy VanderDoes, Gregory Rice, Jack D. Hywood.

Software: Jeremy VanderDoes.

Writing – original draft: Jeremy VanderDoes.

Writing – review & editing: Jeremy VanderDoes, Marie-Liesse Asselin-Labat, Gregory Rice, Jack D. Hywood.

References

1. Toth ZE, Mezey E. Simultaneous visualization of multiple antigens with tyramide signal amplification using antibodies from the same species. Journal of Histochemistry & Cytochemistry. 2007; 55(6):545–554. https://doi.org/10.1369/jhc.6A7134.2007 PMID: 17242468
2. Angelo M, Bendall SC, Finck R, Hale MB, Hitzman C, Borowsky AD, et al. Multiplexed ion beam imaging of human breast tumors. Nature medicine. 2014; 20(4):436–442. https://doi.org/10.1038/nm.3488 PMID: 24584119

3. Giesen C, Wang HA, Schapiro D, Zivanovic N, Jacobs A, Hattendorf B, et al. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. Nature methods. 2014; 11(4):417–422. https://doi.org/10.1038/nmeth.2869 PMID: 24584193

4. Lin JR, Fallahi-Sichani M, Sorger PK. Highly multiplexed imaging of single cells using a high-throughput cyclic immunofluorescence method. Nature communications. 2015; 6(1):8390. https://doi.org/10.1038/ncomms9390 PMID: 26399630

5. Goltsev Y, Samusik N, Kennedy-Darling J, Bhate S, Hale M, Vazquez G, et al. Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. Cell. 2018; 174(4):968–981. https://doi.org/10.1016/j.cell.2018.07.010 PMID: 30078711

6. Herbel C, Dittmer V, Martinez-Osuna M, Kuester LN, Schaefer D, Drewes J, et al. Evaluation of tumor-associated antigen expression with the MACSima high-content imaging platform. Cancer Research. 2019; 79(13_Supplement):4694–4694. https://doi.org/10.1158/1538-7445.AM2019-4694

7. Saka SK, Wang Y, Kishi JY, Zhu A, Zeng Y, Xie W, et al. Immuno-SABER enables highly multiplexed and amplified protein imaging in tissues. Nature biotechnology. 2019; 37(9):1080–1090. https://doi.org/10.1038/s41587-019-0207-y PMID: 31427819

8. Lewis SM, Asselin-Labat ML, Nguyen Q, Berthelet J, Tan X, Wimmer VC, et al. Spatial omics and multiplexed imaging to explore cancer biology. Nature methods. 2021; 18(9):997–1012. https://doi.org/10.1038/s41592-021-01203-6 PMID: 34341583

9. Gaglia G, Burger ML, Ritch CC, Rammos D, Dai Y, Crossland GE, et al. Lymphocyte networks are dynamic cellular communities in the immunoregulatory landscape of lung adenocarcinoma. Cancer cell. 2023; 41(5):871–886.e10. https://doi.org/10.1016/j.ccell.2023.03.015 PMID: 37059105

10. Wang Y, Wang YG, Hu C, Li M, Fan Y, Otter N, et al. Cell graph neural networks enable the digital staging of tumor microenvironment and precise prediction of patient survival in gastric cancer. medRxiv. 2021; p. 2021–09.

11. Lin YWE, Shnitzer T, Talmon R, Villarroel-Espindola F, Desai S, Schalper K, et al. Graph of graphs analysis for multiplexed data with application to imaging mass cytometry. PLoS Computational Biology. 2021; 17(3):e1008741. https://doi.org/10.1371/journal.pcbi.1008741 PMID: 33780435

12. Keren L, Bosse M, Marquez D, Angoshtari R, Jain S, Varma S, et al. A Structured Tumor-Immune Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam Imaging. Cell. 2018; 174(6):1373–1387.e19. https://doi.org/10.1016/j.cell.2018.08.039 PMID: 30193111

13. Keren L, Bosse M, Thompson S, Risom T, Vijayaragavan K, McCaffrey E, et al. MIBI-TOF: A multiplexed imaging platform relates cellular phenotypes and tissue structure. Science advances. 2019; 5 (10):eaax5851–eaax5851. https://doi.org/10.1126/sciadv.aax5851 PMID: 31633026

14. Illian J, Penttinen A, Stoyan H, Stoyan D. Statistical analysis and modelling of spatial point patterns. John Wiley & Sons; 2008.

15. Diggle PJ. Statistical analysis of spatial and spatio-temporal point patterns. CRC press; 2013.

16. Baddeley A, Rubak E, Turner R. Spatial point patterns: methodology and applications with R. CRC press; 2015.

17. Schürch CM, Bhate SS, Barlow GL, Phillips DJ, Noti L, Zlobec I, et al. Coordinated Cellular Neighborhoods Orchestrate Antitumoral Immunity at the Colorectal Cancer Invasive Front. Cell. 2020; 183 (3):838–838. https://doi.org/10.1016/j.cell.2020.10.021 PMID: 33125896

18. Jiang S, Chan CN, Rovira-Clavé X, Chen H, Bai Y, Zhu B, et al. Combined protein and nucleic acid imaging reveals virus-dependent B cell and macrophage immunosuppression of tissue microenvironments. Immunity (Cambridge, Mass). 2022; 55(6):1118–1134.e8. https://doi.org/10.1016/j.immuni.2022.03.020 PMID: 35447093

19. Damond N, Engler S, Zanotelli VRT, Schapiro D, Wasserfall CH, Kusmartseva I, et al. A Map of Human Type 1 Diabetes Progression by Imaging Mass Cytometry. Cell Metabolism. 2019; 29(3):755–768.e5. https://doi.org/10.1016/j.cmet.2018.11.014 PMID: 30713109

20. Jackson HW, Fischer JR, Zanotelli VRT, Ali HR, Mechera R, Soysal SD, et al. The single-cell pathology landscape of breast cancer. Nature (London). 2020; 578(7796):615–620. https://doi.org/10.1038/s41586-019-1876-x PMID: 31959985

21. Färkkilä A, Gulhan DC, Casado J, Jacobson CA, Nguyen H, Kochupurakkal B, et al. Immunogenomic profiling determines responses to combined PARP and PD-1 inhibition in ovarian cancer. Nature communications. 2020; 11(1):1459–13. https://doi.org/10.1038/s41467-020-15315-8 PMID: 32193378

22. Bartlett MS. The Spectral Analysis of Two-Dimensional Point Processes. Biometrika. 1964; 51(3/4):299–311. https://doi.org/10.2307/2334136

23. Ripley BD. The second-order analysis of stationary point processes. Journal of applied probability. 1976; 13(2):255–266. https://doi.org/10.2307/3212829

24. Ripley BD. Modelling Spatial Patterns. Journal of the Royal Statistical Society Series B, Methodological. 1977; 39(2):172–212. https://doi.org/10.1111/j.2517-6161.1977.tb01615.x

25. Carstens JL, Correa de Sampaio P, Yang D, Barua S, Wang H, Rao A, et al. Spatial computation of intratumoral T cells correlates with survival of patients with pancreatic cancer. Nature communications. 2017; 8(1):15095. https://doi.org/10.1038/ncomms15095 PMID: 28447602

26. Barua S, Fang P, Sharma A, Fujimoto J, Wistuba I, Rao AU, et al. Spatial interaction of tumor cells and regulatory T cells correlates with survival in non-small cell lung cancer. Lung Cancer. 2018; 117:73–79. https://doi.org/10.1016/j.lungcan.2018.01.022 PMID: 29409671

27. Bull JA, Macklin PS, Quaiser T, Braun F, Waters SL, Pugh CW, et al. Combining multiple spatial statistics enhances the description of immune cell localisation within tumours. Scientific reports. 2020; 10 (1):18624. https://doi.org/10.1038/s41598-020-75180-9 PMID: 33122646

28. Parra ER, Zhai J, Tamegnon A, Zhou N, Pandurengan RK, Barreto C, et al. Identification of distinct immune landscapes using an automated nine-color multiplex immunofluorescence staining panel and image analysis in paraffin tumor tissues. Scientific reports. 2021; 11(1):1–14. https://doi.org/10.1038/s41598-021-83858-x PMID: 33633208

29. Parra ER. Methods to determine and analyze the cellular spatial distribution extracted from multiplex immunofluorescence data to understand the tumor microenvironment. Frontiers in Molecular Biosciences. 2021; 8:668340. https://doi.org/10.3389/fmolb.2021.668340 PMID: 34179080

30. Canete NP, Iyengar SS, Wilmott JS, Ormerod JT, Harman AN, Patrick E. spicyR: Spatial analysis of in situ cytometry data in R. Health & Medicine Week. 2021; n/a(n/a):7776–.

31. Vu T, Wrobel J, Bitler BG, Schenk EL, Jordan KR, Ghosh D. SPF: a spatial and functional data analytic approach to cell imaging data. PLOS Computational Biology. 2022; 18(6):e1009486. https://doi.org/10.1371/journal.pcbi.1009486 PMID: 35704658

32. Hywood JD, Read MN, Rice G. Statistical analysis of spatially homogeneous dynamic agent-based processes using functional time series analysis. Spatial Statistics. 2016; 17:199–219. https://doi.org/10.1016/j.spasta.2016.06.002

33. Hywood JD, Rice G, Pageon SV, Read MN, Biro M. Detection and characterization of chemotaxis without cell tracking. Journal of the Royal Society Interface. 2021; 18(176):20200879. https://doi.org/10.1098/rsif.2020.0879 PMID: 33715400

34. Seal S, Vu T, Ghosh T, Wrobel J, Ghosh D. DenVar: density-based variation analysis of multiplex imaging data. Bioinformatics Advances. 2022; 2(1):vbac039. https://doi.org/10.1093/bioadv/vbac039 PMID: 36699398

35. Yi M, Zhan T, Peck AR, Hooke JA, Kovatich AJ, Shriver CD, et al. Quantile Index Biomarkers Based on Single-Cell Expression Data. Laboratory Investigation. 2023; 103(8):100158. https://doi.org/10.1016/j.labinv.2023.100158 PMID: 37088463

36. Ramsay JO, Silverman BW. Functional Data Analysis. 2nd ed. Springer series in statistics. New York: Springer; 2005.

37. Barber RF, Candés EJ. Controlling The False Discovery Rate Via Knockoffs. The Annals of statistics. 2015; 43(5):2055–2085. https://doi.org/10.1214/15-AOS1337

38. Bottai G, Raschioni C, Losurdo A, Di Tommaso L, Tinterri C, Torrisi R, et al. An immune stratification reveals a subset of PD-1/LAG-3 double-positive triple-negative breast cancers. Breast cancer research: BCR. 2016; 18(1):121–121. https://doi.org/10.1186/s13058-016-0783-4 PMID: 27912781

39. Lehmann BD, Jovanović B, Chen X, Estrada MV, Johnson KN, Shyr Y, et al. Refinement of Triple-Negative Breast Cancer Molecular Subtypes: Implications for Neoadjuvant Chemotherapy Selection. PloS one. 2016; 11(6):e0157368–e0157368. https://doi.org/10.1371/journal.pone.0157368 PMID: 27310713

40. Sugie T, Sato E, Miyashita M, Yamaguchi R, Sakatani T, Kozuka Y, et al. Multispectral quantitative immunohistochemical analysis of tumor-infiltrating lymphocytes in relation to programmed death-ligand 1 expression in triple-negative breast cancer. Breast cancer (Tokyo, Japan). 2020; 27(4):519–526. https://doi.org/10.1007/s12282-020-01110-2 PMID: 32447649

41. Patwa A, Yamashita R, Long J, Risom T, Angelo M, Keren L, et al. Multiplexed imaging analysis of the tumor-immune microenvironment reveals predictors of outcome in triple-negative breast cancer. Communications biology. 2021; 4(1):852–852. https://doi.org/10.1038/s42003-021-02361-1 PMID: 34244605

42. Greenbaum S, Averbukh I, Soon E, Rizzuto G, Baranski A, Greenwald NF, et al. A spatially resolved timeline of the human maternal-fetal interface. Nature (London). 2023; 619(7970):595–3. https://doi.org/10.1038/s41586-023-06298-9 PMID: 37468587

43. Wang F, Flanagan J, Su N, Wang LC, Bui S, Nielson A, et al. RNAscope: a novel in situ RNA analysis platform for formalin-fixed, paraffin-embedded tissues. The Journal of molecular diagnostics: JMD. 2012; 14(1):22–29. https://doi.org/10.1016/j.jmoldx.2011.08.002 PMID: 22166544

44. Black S, Phillips D, Hickey JW, Kennedy-Darling J, Venkataraaman VG, Samusik N, et al. CODEX multiplexed tissue imaging with DNA-conjugated antibodies. Nature protocols. 2021; 16(8):3802–3835. https://doi.org/10.1038/s41596-021-00556-8 PMID: 34215862

45. Chen HY, Palendira U, Feng CG. Navigating the cellular landscape in tissue: Recent advances in defining the pathogenesis of human disease. Computational and structural biotechnology journal. 2022; 20:5256–5263. https://doi.org/10.1016/j.csbj.2022.09.005 PMID: 36212528

46. Williams CG, Lee HJ, Asatsuma T, Vento-Tormo R, Haque A. An introduction to spatial transcriptomics for biomedical research. Genome medicine. 2022; 14(1):1–68. https://doi.org/10.1186/s13073-022-01075-1 PMID: 35761361

47. Li Q, Ren Z, Cao K, Li MM, Wang K, Zhou Y. CancerVar: An artificial intelligence-empowered platform for clinical interpretation of somatic mutations in cancer. Science advances. 2022; 8(18):eabj1624–eabj1624. https://doi.org/10.1126/sciadv.abj1624 PMID: 35544644

48. Canning AJ, Viggiano S, Fernandez-Zapico ME, Cosgrove MS. Parallel functional annotation of cancer-associated missense mutations in histone methyltransferases. Scientific reports. 2022; 12 (1):18487–18487. https://doi.org/10.1038/s41598-022-23229-2 PMID: 36323913

49. Sadeghirad H, Liu N, Monkman J, Ma N, Cheikh BB, Jhaveri N, et al. Compartmentalized spatial profiling of the tumor microenvironment in head and neck squamous cell carcinoma identifies immune checkpoint molecules and tumor necrosis factor receptor superfamily members as biomarkers of response to immunotherapy. Frontiers in immunology. 2023; 14:1135489–1135489. https://doi.org/10.3389/fimmu.2023.1135489 PMID: 37153589

50. Zanotelli VR, Leutenegger M, Lun XK, Georgi F, de Souza N, Bodenmiller B. A quantitative analysis of the interplay of environment, neighborhood, and cell state in 3D spheroids. Molecular Systems Biology. 2020; 16(12):e9798. https://doi.org/10.15252/msb.20209798 PMID: 33369114

51. R Core Team. R: A Language and Environment for Statistical Computing; 2022. Available from: https://www.R-project.org/.

52. Breiman L. Random Forests. Machine learning. 2001; 45(1):5–32. https://doi.org/10.1023/A:1010933404324

53. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and Regression Trees. Chapman and Hall/CRC; 1984.

54. Therneau T, Atkinson B. rpart: Recursive Partitioning and Regression Trees; 2022. Available from: https://CRAN.R-project.org/package=rpart.

55. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC bioinformatics. 2007; 8(1):25–25. https://doi.org/10.1186/1471-2105-8-25 PMID: 17254353

56. Thomas M. A Generalization Of Poisson's Binomial Limit For Use In Ecology. Biometrika. 1949; 36(1-2):18–25. https://doi.org/10.1093/biomet/36.1-2.18 PMID: 18146215

57. Baddeley AJ, Møller J, Waagepetersen R. Non-and semi-parametric estimation of interaction in inhomogeneous point patterns. Statistica Neerlandica. 2000; 54(3):329–350. https://doi.org/10.1111/1467-9574.00144

58. Boots BN, Getis A. Point pattern analysis. Scientific geography series; v. 8. Newbury Park, Calif: Sage Publications; 1988.

59. Zhu X, Zhang J, Xu Y, Wang J, Peng X, Li HD. Single-Cell Clustering Based on Shared Nearest Neighbor and Graph Partitioning. Interdisciplinary sciences: computational life sciences. 2020; 12(2):117–130. PMID: 32086753

60. Penttinen A, Stoyan D, Henttonen HM. Marked point processes in forest statistics. Forest science. 1992; 38(4):806–824. https://doi.org/10.1093/forestscience/38.4.806

61. Ishwaran H, Kogalur UB, Chen X, Minn AJ. Random survival forests for high-dimensional data. Statistical analysis and data mining. 2011; 4(1):115–132. https://doi.org/10.1002/sam.10103

62. Pickett KL, Suresh K, Campbell KR, Davis S, Juarez-Colunga E. Random survival forests for dynamic predictions of a time-to-event outcome using a longitudinal biomarker. BMC medical research methodology. 2021; 21(1):1–216. https://doi.org/10.1186/s12874-021-01375-x PMID: 34657597