Quantitative Spatial Analysis of Organotropism Using Zebrafish Models

Hyeyun Christine Lee

Received September 04, 2024
Accepted January 14, 2025
Electronic access January 31, 2025

Cancer metastasis, the spread of cancer cells from primary tumor sites to distant organs, remains the leading cause of cancer-related deaths. Organotropism, the selective tendency of cancer cells to metastasize to specific organs, impacts clinical outcomes as it affects therapeutic responses, which can vary widely depending on the affected organ. However, studying organotropism has been challenging due to weak spatial correlations and limitations in tracking metastatic behavior across individual patients. This study develops a computational pipeline for quantifying organotropic behavior of cancer cells using spatial statistics metrics in a zebrafish xenograft model. From a previously established workflow, metastatic sites from high-resolution imaging data were transformed into point patterns. Spatial statistical metrics were employed to quantify global and local heterogeneity of point patterns representing metastatic sites. To validate the pipeline, we employed the Fish Point Pattern Simulator (FPPS) to generate point patterns that mimic zebrafish models with varying metastatic heterogeneity. The validated pipeline was applied to TC32 Ewing sarcoma, relevant to organotropic mechanisms, and fibroblast cells, which do not exhibit such tendencies. The results identified significant differences in their global and local spatial point pattern distributions, suggesting differential metastatic profiles. This highlights the pipeline's efficacy in quantifying metastatic propensity and comparing the behavior of different cell lines.

Introduction

Cancer cells display remarkable adaptability, enabling them to survive and proliferate in diverse and often adverse environments. The spreading of cancer cells from the primary site to distant organs, cancer metastasis, remains a major challenge in oncology as the leading cause of cancer-related deaths[1]. This process involves intricate interactions between cancer cells and their microenvironments, resulting in the formation of secondary tumors in biased and specific– rather than random– organs[2]. This is illustrated in Figure 1A. This is a phenomenon known as organotropism, which is the attraction of cells to a particular organ or tissue of the body during metastasis2. This differential colonization is clinically significant, as metastases in distinct organs often exhibit varied responses to therapies, impacting treatment outcomes[3,4]. For pediatric cancers like Ewing sarcoma, which are often characterized by a single genetic mutation, such differential survival rates are more likely attributable to environmental factors than to genomic heterogeneity[5]. To study organotropism, researchers have employed various methods, including organoid models and imaging techniques. In a prior study, Saucier et al. utilized the zebrafish xenograft imaging method to visualize metastatic cancer cells spreading patterns within zebrafish larvae[6]. These are a valuable model for studying human cancers due to their rapid development, optical transparency, and ability to model various human tumor types with similar morphology and signaling pathways[7]. Zebrafish also share approximately 70% of human genes and exhibit functional similarities in organs like the liver and kidney, making them an excellent model for exploring organ-specific interactions in metastasis[8]. After imaging, this study employed advanced image processing techniques, to minimize the effect of noise and detect hotspots of metastatic colonization in the fish[9]. Despite advancements in imaging, computationally comparing the spatial distribution of xenografted cancer cells remains a significant challenge. Irregular shapes, diverse metastatic sites, and the need for robust statistical validation complicate the process. Specifically in zebrafish models, irregular edges, sensitivity of many spatial statistics metrics at the edges, and weak spatial patterns present challenges in distinguishing results from what is known as "meaningful" heterogeneity vs. "meaningless" (due to noise) heterogeneity. This project addresses this gap by converting imaged metastatic sites in zebrafish (visualized as regions in Fig. 1A) into point-based spatial distributions (Fig. 1B) to represent the sites of metastasis; this spatial distribution gives us the statistical power to quantify and compare metastatic propensity. This study aims to enhance the understanding of cancer organotropism by comparing spatial statistical measures that can provide a robust framework for comprehensively quantifying and computationally comparing metastatic profiles from imaged xenograft data. For this objective, the hypothesis

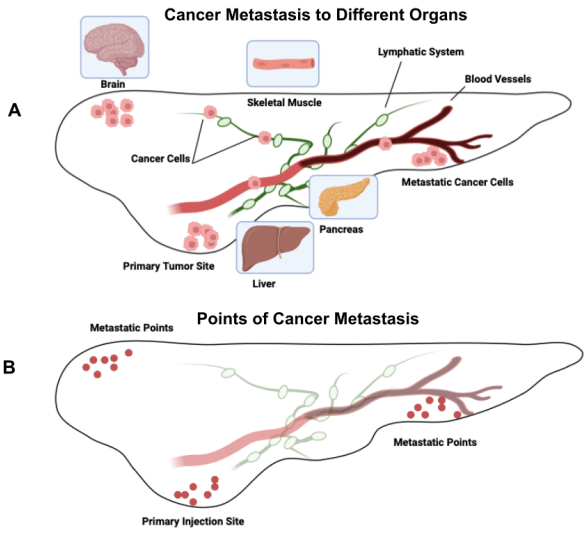

Fig. 1 (A) Cancer metastasis pathway through the lymph and blood system to different organs on a zebrafish template. (B) Newly colonized cells are detected as points (see Methods) in varying organs.

that a computational framework can quantitatively measure and compare different metastatic profiles of cell lines injected into zebrafish models guides our research. This approach introduces global and local computational metrics to assess spatial heterogeneity, providing a novel framework for quantifying and comparing metastatic patterns.

To validate these metrics, a Python-based FPPS was developed to generate spatial distributions based on user-defined parameters[10]. After confirming the reliability of the metrics, the spatial profiles of TC32 Ewing sarcoma and fibroblast cells were analyzed using unpublished xenograft imaging data. The TC32 Ewing sarcoma and fibroblast cell lines are ideal for assessing spatial statistical models because they present contrasting behaviors: while TC32 cells exemplify the organotropic tendencies of pediatric cancers like Ewing sarcoma, non-transformed fibroblast cell lines lack such behavior. This distinction allows an evaluation of the metastatic profile of TC32 and fibroblast cells to exhibit a controlled analysis of the behavior of these computational statistical models in detecting meaningful heterogeneity in different cell lines spatial distributions.

Results

Global Spatial Analysis

This study developed computational frameworks to assess the spatial heterogeneity of metastatic sites of cancer cells in a zebrafish model. To assess the similarity metastatic profiles, metastatic sites of the injected cells were represented by a point pattern in a probabilistic distribution in a zebrafish template. We employed various computational statistical metrics to compare these probabilistic distributions[11]. First, the Kullback-Leibler's Divergence (KLD), an information-based metric, measured the divergence between two probability distributions (p, q) by determining the degree of information lost when q assumes p[12,13]. To validate the effectiveness of the KLD in measuring the spreading locations of metastatic hotspots in zebrafish, we generated simulated point patterns in a 2D fish template using a Python-based package, FPPS. These patterns were modeled by using a Gaussian distribution for the colonization sites of cancer in a fish, while the movement of single cells throughout the fish's body was represented by a uniform random distribution. For analysis, these discrete points were converted into a continuous distribution using Kernel Density Estimation maps (see Methods). Initially, we examined the increasing divergence of overlapped metastatic hotspots between two patterns. At the beginning the metastatic hotspot in Distribution A and Distribution B were completely overlapped. Through multiple trials, the center of the hotspots in Distribution B were shifted away from that in Distribution A. To estimate the variability of the KLD in similar hotspots with random variation, we regenerated the distribution with the same parameters using five replicants. As shown in the graph, the divergence between Distribution A and Distribution B increases as the center of the metastatic hotspots in Distribution B shifts (Table 1), until they are eventually fully separated (Fig. 2A). There is a large variation in the error bars for small dX values. The statistical significance of the divergence was determined using a bootstrapping method to calculate permutation p-values (see Methods). As linear heterogeneity (dX) increases between the distributions, the p-values decrease below 0.05, representing that the distributions are statistically different. By using the probabilistic KLD, we quantified the overlap, or similarity, between spreading patterns of two distributions, considering the intensity and location of hotspots in the probability distribution. When the hotspots of Distribution B are sufficiently separated from those of Distribution A, the KLD value remains relatively constant (Table 1). This means that the KLD exhibits a saturation behavior for dX values greater than 20. As a result, the exact distance between the hotspots of these distributions becomes less significant and is not fully captured by this metric.

We then used a cost-based metric, the Earth Mover's Distance (EMD) metric, to measure the divergence between two probability distributions (p, q) by determining the minimum cost of the transformation to turn one distribution into the other (see Methods)[14]. We used the same simulated point patterns, mentioned above, to validate the effectiveness of the EMD compared to the KLD. The EMD considers the differences in the spatial location and arrangement of the points between distributions, effectively capturing the relative positioning and clustering of hotspots. As the heterogeneity increases in the mean of the two hotspots, the EMD increases

in a constant manner (Table 2). As shown in the EMD graph, when the clustered regions of Distribution B are increasingly separated from Distribution A, the EMD increases, and the corresponding p-value decreases, highlighting differences between the distributions (Fig. 1D, Table 2). Unlike the KLD, the EMD sensitively measures the distance between hotspot regions of the two distributions past their overlap, where the KLD becomes saturated, making it more effective for evaluating the dissimilarity between spreading patterns. As a cost-of-transformation metric, the EMD offers distinct advantages over the KLD in detecting differences between two overlapped distributions. Based on our results, both the KLD and the EMD evaluate the metastatic profiles of cancer cell lines in zebrafish at a global scale.

Next, we tested the behavior of these computational statistical methods in examining the divergence of metastatic hotspots of varying sizes, with and without overlap. We increased the size of the hotspot of Distribution B, the sigma parameter in the FPPS, while holding that of the first distribution constant. Our results show that both the KLD and EMD capture differences in hotspot size. While the KLD increases with the hotspot size, it eventually saturates when the hotspots in the two distributions are sufficiently separated; at high Sigma Change values greater than 25, there is a large variation in error bars and saturation behavior in the metric (Table 3). In contrast, the EMD continues to increase monotonically as the hotspot expands in Distribution B, reflecting its sensitivity to the spatial rearrangement between these distributions rather than merely considering the overlap (Table 4).

After validating the following global metrics, the established computational frameworks were applied to unpublished xenograft imaging data to assess their effectiveness in quantitatively comparing the metastatic profile of two different cell lines. The data was collected by injecting TC32 Ewing sarcoma and NIH 3T3 fibroblast cells into zebrafish embryos and tracking their metastatic spread using advanced image processing techniques. The imaged points were mapped onto an x,y plane, and spatial statistical measures were applied to analyze the resulting point clouds, creating a distribution map (see Methods).

The p-value of the KLD between Fibroblast and TC32 is greater than 0.05 (Fig. 3C). However, as the simulator experiment validated, it is unsure if the high p-value is due to the lack of meaningful heterogeneity in metastatic hotspot locations or the saturating nature of the KLD metric. For the EMD, the p-value is less than 0.05 showing that the two distributions are significantly different (Fig. 3C). The linear nature of the EMD and the data from the simulator render it a reliable measure of heterogeneity between the Fibroblast and TC32 spatial data on the zebrafish xenograft. This starts to reveal that this computational framework is effective in highlighting that there is a significant difference between the metastatic propensity of fibroblast and that of TC32. Though it is likely there is meaningful heterogeneity between these two distributions on a global scale, the questions of how to analyze the number, location, and correlations of different hotspots on the local distribution remain. Further local spatial analysis was conducted to examine these questions.

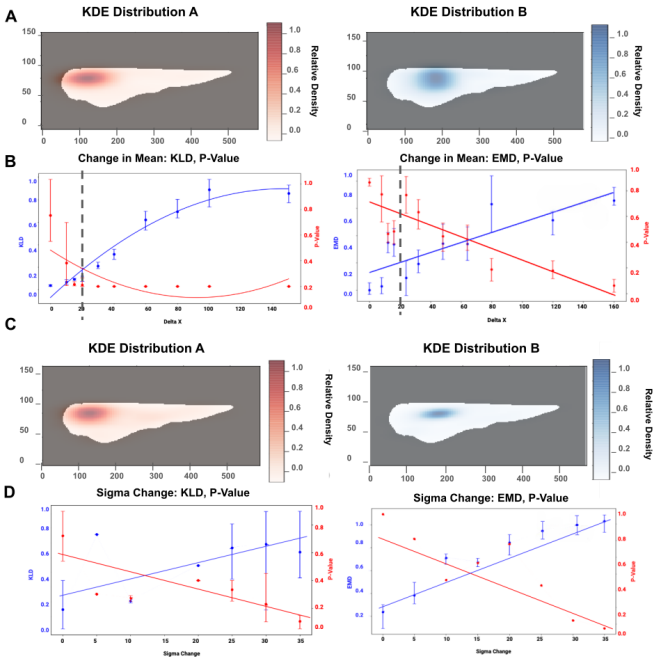

Fig. 2 (A) Fish point pattern simulator with the following parameters: Mean A (100,100), Mean B: (180, 100), Delta X: 80. A kernel density estimate is mapped onto each distribution of simulated points. (B) Graph of KLD, EMD, and respective p-values as the mean of Distribution A is held constant and that of Distribution B is moved further to the right (Delta X increases). The dotted line shows the Delta X value where the two hotspots stop overlapping. (C) Fish point pattern simulator with the following parameters: Mean A (100,100), Mean B: (150, 100), Sigma Change: 5. A kernel density estimate is mapped onto each distribution of simulated points. (D) Graph of KLD, EMD, and respective p-values as the sigma of Distribution B increases (Sigma Change increases). $P < 0.05$ detects statically significant heterogeneity between the distributions.

Local Spatial Analysis

Though the above global measures of spatial dissimilarity provide a framework to compare the overall metastatic profiles, additional methods are needed to reveal specific locations of dissimilarities, identify the number of hotspots, and compare hotspot locations. To reveal local spatial patterns in the cell spreading distribution, the Mean Shift Clustering (MSC) and Distance to Reference Point (DRP) metrics were employed. The MSC metric uses an iterative approach to find the number, location, and size of hotspots in the distribution (see Methods)[?].

Table 1: KLD Centroid Shift (dX)

dX	0	5	10	15	20	30	40	60	80	100	150
P-Value	0.92, 0.80, 0.83, 0.37, 0.30	0.37, 0.70, 0.83, 0.76, 0.41	0.26, 0.39, 0.23, 0.29, 0.31	0.26, 0.39, 0.23, 0.29, 0.31	0.15, 0.08, 0.11, 0.19, 0.19	0.05, 0.10, 0.18, 0.17, 0.04	0.15, 0.10, 0.04, 0.13, 0.12	0.01, 0.04, 0.07, 0.14, 0.16	0.01, 0.02, 0.01, 0.00, 0.03	0.03, 0.02, 0.02, 0.01, 0.01	0.00, 0.00, 0.01, 0.02, 0.03

Table 2: EMD Centroid Shift (dX)

dX	0	5	10	15	20	30	40	60	80	100	150
P-Value	0.93, 0.87, 0.83, 0.97, 0.93	0.97, 0.50, 0.83, 0.76, 0.48	0.36, 0.39, 0.53, 0.59, 0.31	0.44, 0.47, 0.46, 0.50, 0.37	0.65, 0.88, 0.51, 0.49, 0.69	0.75, 0.54, 0.58, 0.67, 0.54	0.35, 0.31, 0.54, 0.43, 0.42	0.31, 0.34, 0.36, 0.24, 0.46	0.11, 0.12, 0.31, 0.10, 0.13	0.02, 0.03, 0.02, 0.01, 0.01	0.04, 0.00, 0.01, 0.02, 0.03

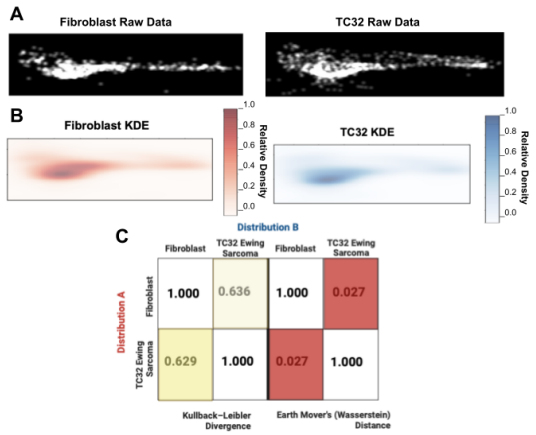

Fig. 3 (A) Processed images of Fibroblast and TC32 cells in zebrafish xenografts. Detected points of the fibroblast and the TC32 cells are each accumulated into an accumulator (n=65). (B) A Kernel Density Estimate map is created for each distribution. (C) The discrimination matrix shows the results of the KLD, EMD, and the corresponding p-values between the fibroblast and TC32 distribution.

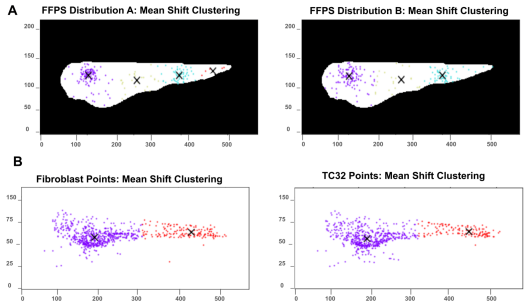

Fig. 4 (A) Simulated points on the FPPS. The X represents the mean of each Mean Shift Clustering detected hotspots. The color of the point represents the hotspot it was assigned to by Mean Shift Clustering. The MSC mean and hotspot sizes were compared to the input FPPS values and the percent error was calculated. (B) Results of the Fibroblast (Distribution 1) and TC32 (Distribution 2) points Mean Shift Clustering. The X represents the detected means and each color represents the hotspot the points were assigned to. Hotspot 1A Mean = (118.11, 87.34), 1A Size = 603, 1B Mean = (389.16, 68.88), 1B Size = 161, 2A Mean = (118.07, 89.90), 2A Size = 562, 2B Mean = (420.10, 64.07), 2B Size = 129.

To validate the effectiveness in quantifying the local pattern of point cloud behavior, specific metastatic clustering profiles were simulated. The parameters were compared to the outputs of the MSC (see Table 5).

As shown in the figure, the MSC accurately detected the number of metastatic hotspots and the (x,y) centers of these hotspots; it also used an iterative approach to assign cells into the appropriate hotspots, showing which cells belong to which metastatic hotspot. Each hotspot created by the FPPS parameters were detected by the MSC method within a 5% error for the x and y location of the hotspot center (Fig. 4A); as the percent error between the centers of the FPPS simulated centers and the MSC detected centers are less than 5%, it can be attributed to noise in the metric (see Table 5 in Methods). This computational framework of local spatial analysis starts to reveal localized patterns in the point cloud with multiple metastatic hotspots despite single-cell noise on a zebrafish.

After validating this metric, the MSC framework was applied to assess its ability to detect metastatic hotspots in the Fibroblast and TC32 data. Two different hotspots were detected in both of the point clouds, revealing that there are two significant metastatic hotspots in the point cloud patterns where cells preferentially colonized and formed metastatic groups (Fig. 4B). The significance of the hotspots detected by the MSC highlights that the cells were spreading due to some underlying biological phenomena rather than random dispersal or dispersal due to growth. The mean of the first (left) hotspot is similar for both of the point clouds as the percent differences in their x and y locations of the hotspot center is less than 5% (0.002% and 2.930% respectively). However, the mean of the second (right) hotspot near the tail is different as the percent difference in their center x and y locations are more than 5% (7.38% and 7.51% respectively); the mean of hotspot 2B in the TC32 distribution is significantly further right towards the end of the tail compared to 1B in the Fibroblast distribution (see Table 6 in Methods) This local analysis raises the possibility of localized dissimilarities in the metastatic propensity near the tail, as the centroid and size of each hotspots can be examined separately in the context of an x,y plane, rather than just the general dispersal of cells using global

Table 3: KLD Sigma Change

Sigma Change	0	5	10	20	25	30	35
P-Value	0.95, 0.89, 0.99, 0.98, 0.96	0.97, 0.50, 0.83, 0.76, 0.48	0.8, 0.79, 0.82, 0.79, 0.83	0.45, 0.45, 0.50, 0.52, 0.49	0.75, 0.69, 0.70, 0.78, 0.69	0.09, 0.10, 0.04, 0.30, 0.12	0.01, 0.02, 0.01, 0.05, 0.03

Table 4: EMD Sigma Change

Sigma Change	0	5	10	20	25	30	35
P-Value	0.9, 0.85, 0.54, 0.6, 0.76	0.3, 0.28, 0.33, 0.29, 0.3	0.29, 0.25, 0.25, 0.3, 0.22	0.40, 0.35, 0.35, 0.34, 0.39	0.35, 0.33, 0.40, 0.25, 0.21	0.06, 0.10, 0.04, 0.29, 0.04	0.03, 0.02, 0.05, 0.01, 0.03

distribution metrics. This computational analysis mapping of the local hotspots on an x,y plane provides the framework for a future cross-study of the centroids of the means and the mapped microenvironments of zebrafish on the same x,y plane. The scope of this study was to create quantitative metrics to measure spatial heterogeneity, and therefore, the exploration of the localized hotspots in relation to specific microenvironments was not further pursued here but remains an area for future research.

After detecting and analyzing the size and locations of local hotspots with the MSC, the DRP metric was employed to study the correlation and relationships between these hotspots; the points in each hotspot were analyzed in relation to the injection site, the perivitelline space (PVS), as the distance that the cells traveled from the primary tumor site. A constant reference point was set near the injection site in the PVS and the distance was measured as d. The distances were graphed and investigated for correlation, or the lack thereof, between hotspots in different metastatic profiles (see Methods).

Again, points modeling metastatic profiles with multiple hotspots were generated by providing multiple parameters for the mean of the Gaussian clusters. Multiple Gaussian clusters modeled the multiple sites of metastatic colonization in the zebrafish model, representing the tendency of cell lines to colonize different organs, often with similar microenvironments [15]. One cluster was generated with the same parameters in both distributions to model similar metastatic colonization in the same location; another cluster was generated with a significantly different mean parameter to model a metastatic hotspot that is only found in one specific cell line and not the other (Fig. 5A). From the t-test, a p-value less than 0.05 shows that two clusters are significantly different between the distributions. As seen in the graph, the first peak is significantly indifferent in the blue and red distributions, showing that the specific metastatic cluster is similar between both distributions. On the other hand, the second peak is significantly different between the blue and red distributions as the blue peak is further to the right of the graph, confirming that the metastatic hotspot is a greater distance from the primary injection site (Fig. 5B).

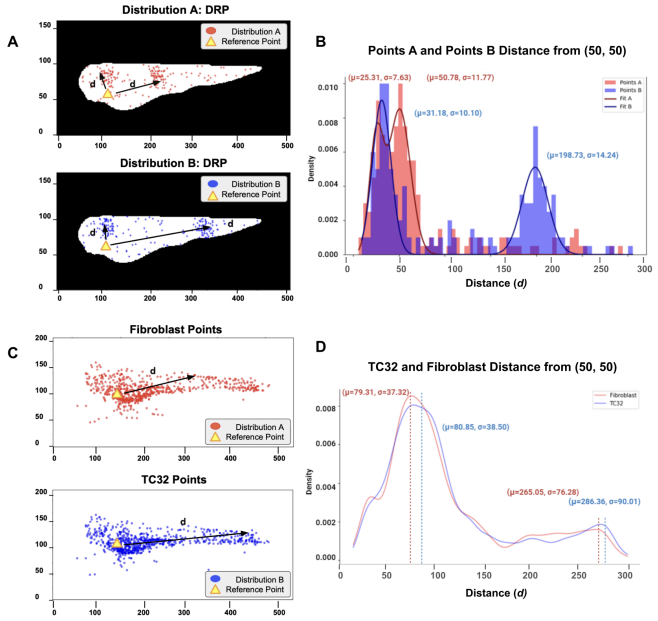

Fig. 5 (A) FPPS generated points with two hotspots. While both of the hotspots had the same sigma (10) for both distributions, Cluster 1 had the same mean for both distributions (100, 100) and Cluster 2 had different means for Distribution A (200, 100) and Distribution B (350, 100). The distance between points in the distribution and the reference point was measured as d. (B) Histogram of d values for Distribution A and Distribution B. A Multi Gaussian curve is fit over the histogram and the resulting mu and sigma is noted above the corresponding peak. The p-value for the first peak is 8.48e-01 and the p-value for the second peak is 5.11e-28. (C) Points of fibroblast and TC32 cells with the reference point. d is measured from the reference point near the injection site and all of the points in each distribution. (D) Multi Gaussian fit curve over the distances to the reference point. The mu and sigma of each peak is labeled with the corresponding colors. The p-value of the t-test is 6.11e-01 and 2.67e-01 for the first and second hotspots respectively.

After validating the DRP metric, it was applied to the Fibroblast and TC32 point pattern data to assess its ability in quantifying the correlation of different hotspots in different cell lines. The Multi-Gaussian fitting of the DRP data detected two clusters in both point clouds meaning that both distributions have two distinct metastatic hotspots (Fig. 5D); this supported the findings from the MSC metric. The DRP metric reveals additional information about the correlation between these multiple clusters in the two distributions. The local metric allows for the division of the metastatic profile into local clusters that can individually be analyzed for correlation, size, and location. Cluster 1 was detected as statistically indifferent ($p > 0.05$) while Cluster 2 was detected as statistically dissimilar ($p < 0.05$) between the distributions. This, in addition to the MSC, shows there are two distinct metastatic clusters in both distributions and the first cluster is indifferent while the second cluster (near the tail) is significantly different between the two point clouds. This computational pipeline provides a framework to effectively analyze localized patterns of hotspots in the metastatic profile of different cell lines.

Discussion

This study established a robust computational framework for analyzing point patterns representing sites of cancer cell metastasis, at both a global and local scale, to quantify the metastatic propensity of cancer cells in a zebrafish model. We applied KLD and EMD to assess global similarities between point distributions, though the KLD exhibited a saturating behavior at distributions with high linear heterogeneity. Both global metrics used in this study offer advantages over the Jaccard index, used in previous studies, by measuring heterogeneity without relying on overlap and incorporating cluster intensity through probability distributions. Though these global metrics couldn't detect local spatial patterns, the MSC identified the number, concentration, size, and location of cell clusters; the DRP metric analyzed the spatial arrangement relative to the injection site in ways that weren't possible previously. A t-test quantified the significance of the correlation, isolated from noise, between clusters in each distribution, providing insights into the metastatic propensity of different cells.

After applying the established computational framework to unpublished data, the results indicated that the framework can detect that TC32 Ewing sarcoma cells exhibit statistically different global spatial distributions compared to fibroblast cells as shown by EMD. In contrast, KLD analysis did not reveal significant differences, though our FPPS data reveals it is likely due to the probabilistic nature of the metric that causes saturation of the metric at high degrees of dissimilarity. The distinct global distribution patterns or metastatic point clouds highlights that TC32 cells have a non-random differing success rates of proliferation in various parts of the zebrafish compared to fibroblast cells, potentially influenced by factors such as differential adhesion to organ-specific extracellular matrices or responsiveness to local growth factors in different microenvironments. These biological phenomena could underlie the observed disparities in metastatic behavior. The MSC, a local measure of spatial data, detected two clusters for both the TC32 and fibroblast cells. These results of two distinct, yet differing clusters highlight that favorable conditions, such as local microenvironmental factors or cell-to-microenvironment interactions, potentially enables cells to cluster and proliferate in these hotspots. Building onto this, the DRP metric highlighted heterogeneity in metastatic colonies in the tail; the statistically different locations of the second cluster, confirmed by the t-test p-value less than 0.05, raise the potential that a non-random biological phenomenon underlies the metastasis and proliferation of TC32 cells to the tail. This guides future research regarding the biological phenomena, i.e. potentially favorable cancer interactions with the hypoxic tail microenvironment, underlying the spatial distributions revealed in this quantitative statistical study. These findings emphasize the value of considering both global and local patterns in metastatic behavior, as each reveals different important information about the metastatic propensity [16].

However, our research has limitations. As mentioned, KLD metric showed saturation and reduced sensitivity at high divergence levels, potentially affecting its reliability in distinguishing between highly heterogeneous distributions. Further analysis would be needed to determine whether the insignificant p-value for KLD between TC32 and fibroblast point clouds is due to true statistical indifference or the metric's saturation issue. Also, the DRP metric, which analyzes distances in a circular (radius) manner, may not fully capture the complexities of metastatic spread in three-dimensions. For example, two equidistant clusters but on opposite sides of the injection site could be incorrectly grouped as a single cluster by the DRP. Finally, the zebrafish model, while valuable models for studying metastasis, may not entirely replicate the human cancer metastasis process, potentially limiting the generalizability of our findings. Key differences, such as the lack of a fully developed adaptive immune system in zebrafish larvae and variations in tissue architecture and extracellular matrix composition, can influence tumor progression and microenvironment interactions, potentially affecting the generalizability of the findings to human biology. However, the zebrafish model's advantages, including real-time visualization of metastatic behavior, genetic similarity to humans, and high-throughput capability, often outweigh these limitations, providing crucial insights into the mechanisms of metastasis that can guide further validation. Though the scope of this study establishes computational statistical measures to compare point patterns representing metastatic profiles, future

research should focus on translating zebrafish model results to human organs. Understanding the correlation between zebrafish and human microenvironments could provide insights into the factors driving specific metastatic profiles.

Other areas of future research include using the established pipelines to measure the metastatic distributions of different cell lines. These metrics enable scientists to explore differences in point cloud distributions representing metastatic spreading patterns under various experimental conditions and drug treatments.

Additionally, the microenvironments in the location of clusters detected by the metrics and their effects on cell proliferation and metastasis can be examined in future research. Knowing where specific cancer lines differentially locate can direct researchers to study the unique characteristics of these microenvironments, such as hypoxia, vascular density, or specific growth factor levels, that support cancer cell survival and growth [17]. For example, observing cancer cell proliferation in areas rich in specific growth factors can highlight pathways for interrupting angiogenesis or immune evasion . By comparing the microenvironments associated with different clusters, researchers can identify specific factors influencing metastatic behavior. This knowledge could guide the development of targeted drugs and treatments aimed at these microenvironments. Finally, after a comprehensive accumulation of a zebrafish tissue map is established, organ-specific markers during the imaging process could further study if the heterogeneity in the metastatic distributions are in the same or different organs in the zebrafish. This future research would help solidify the exact organ or tissue destination of the injected cancer cells.

Methods

To analyze the spatial distribution of cancer cells, the imaged points were mapped onto an x,y plane and a Kernel Density Estimate (KDE) algorithm was applied. The KDE, optimized using Scott's Rule, created a continuous probability density distribution, smoothing the data points and revealing areas of higher concentration. This method avoids the sharp, discontinuous boundaries often seen with histograms, allowing for clearer visualization and identification of clusters [18]. Finally, we employed various statistical metrics using Python-based pipelines, including KLD, EMD, MSC, and DRP Multi Gaussian fit, to compare metastatic profiles. These metrics were categorized into global and local spatial analyses. The KLD measures the difference between two probability distributions, providing insights into the variability of cell distributions on the zebrafish. The KLD is measured as the log-loss of the expected information lost when the distribution Q is assumed in distribution P. At each location, the KDE probability density was used to compare the divergence between the two probability distributions.

$$KLD(P\|Q) = \sum_i P(i) \log\left(\frac{P(i)}{Q(i)}\right)$$

The KLD is directional meaning that distribution Q is compared relative to distribution P. For the analysis between TC32 and Fibroblast data, the KLD was considered in both directions and recorded in the discrimination matrix (see Fig. 3). The EMD, also known as Wasserstein Distance, is a cost-of-transformation metric that quantifies the difference between two distributions. It calculates the minimum amount of work required to transform one distribution into the other by considering both the mass to be moved and the Euclidean distance over which it must be transported; this quantifies the difference in cell spreading by considering how far and how much of the cell mass migrated to different colonization sites on a global scale.

$$\text{EMD}(P\|Q) = \min \sum_i \sum_j M_{ij} d_{ij}$$

To determine if the output KLD and EMD are significant, meaning it detects significant heterogeneity, a permutation and bootstrapping analysis was conducted to calculate the p-value. The accumulated points from Distribution A and Distribution B are combined into one Joint Null distribution. The points from the Joint Null distribution are randomly sampled into Distribution A' and Distribution B' to create null accumulators for K=1,000 iterations, representing the KLD and EMD values in the presence of random, insignificant variation. The p-value is the area of the distribution greater than the experimental value divided by the total area for the Jaccard and the area of the distribution less than the experimental value divided by the total area for the divergence metrics (KLD and EMD). A p-value less than 0.05 confirmed significantly meaningful heterogeneity between the distributions, not due to random chance.

For local spatial analysis, the MSC identifies and assigns clusters of cancer cells based on their spatial distribution. Unlike K-Means clustering, MSC does not require an initial input of the number of clusters. Instead, each point in the dataset is moved iteratively towards areas of higher density by computing the mean of the points within a defined window (kernel) centered on the current point, and then shifting the points towards this mean. This process continues for a maximum of 300 iterations until the mean points converge to locations of local maximum. After convergence, the points are assigned to clusters based on their proximity to the nearest mean after shifting. We then compared the cluster means (x, y), cluster counts, and cluster sizes between distributions. The percent difference between the corresponding clusters (i.e. Cluster 1A vs. Cluster 1B) was calculated. A percent error greater than the percent difference from the FPPS MSC considered significant, indicating meaningful heterogeneity rather than due to random variation (Table 2).

Building on the local analysis of metastatic cell clustering , the DRP method was employed to start to see local spatial patterns in relation to the injection site. This method measures the distance (d) of each point to a reference point set near the injection, providing a spatial context for how cells metastasize from the origin. The distances were graphed on a normalized histogram for both distributions, and a Multigaussian curve was fitted over the histogram to see the mean () and the standard deviation () of the peaks. Based on Euclidean distances, the points are assigned into clusters for correlation analysis. A t-test was conducted to compare the peaks of the Multigaussian curves, assessing the correlation between clusters in the metastatic profile. A p-value less than 0.05 was considered to be statistically significant, indicating that the clusters were different or uncorrelated.

After identifying the metrics to be utilized, we needed to validate the accuracy, behavior, and output. To achieve this, a FPPS was created to test the metrics on simulated fish point data. A zebrafish template was used to create a binary mask on an x,y grid, within which points were generated according to specific parameters: the number of points, the number of clusters, the number of uniform points, the proportion of the total number of points in each cluster, the location (x, y) of each cluster's mean, and the sigma of each cluster. This model simulated the organotrophic behavior of metastatic cells with multiple Gaussian clusters and the noise found in raw experimental data with uniform points. After generating and analyzing the simulated points, a KDE map was created using the same algorithm applied to the experimental data. The methods used for permutation tests were then applied to the simulated data to ensure consistency in analysis.

To validate the two global metrics, KLD and EMD, the FPPS generated two types of heterogeneity: linear, where the mean of the clusters changes, and radial, where the sigma of the clusters changes It was noted if the detection of heterogeneity continues past the overlap of clusters, different from the Jaccard. For linear heterogeneity, points for Distribution A and Distribution B were generated with the following initial parameters: Mean A = (100,100), Mean B = (180, 100), Sigma = 10. Then, Mean A was held constant while Mean B was moved to the right in intervals of 10, with the sigma of both clusters held at 10. This was repeated five times for each parameter, and error bars were added to the graph to represent the metric's sensitivity to random variations. The KLD and EMD were measured and the corresponding p-value was recorded, and a curve of best fit was applied to the graph to visualize the general trend of the metric as the degree of heterogeneity increases. For radial heterogeneity, the initial parameters were Mean A = (100,100), Mean B = (150, 100), Sigma A = 10, and Sigma B = 0. Sigma B was then increased in intervals of 1 while all of the other parameters were held constant. This process was also repeated five times, with the error bars as above. The KLD and EMD were measured again, and p-value was recorded. In both experiments, a p-value less than 0.05 was significant, indicating the global distributions were statistically different.

To validate the two local metrics, the FPPS generated multiple clusters with background noise. The metrics were assessed based on their ability to detect the number of clusters, the relative location of the different clusters, and the correlation between clusters.

First to validate the MSC metric, points for Distribution A and Distribution B were generated with the specific parameters (see Table 5). The MSC results were then compared to the input parameters, and the mean percent error was calculated. The mean percent error was determined separately for the x values and the y values, while the sigma percent error was calculated between clusters. This error represents the variation in MSC results due to non-statistical heterogeneity, such as noise or algorithm limitations. As the table shows, the results of the TC32 and Fibroblast MSC included the number of clusters and the mean and sigma of each cluster. To assess the statistically-significant similarity between the clusters, the percent difference was calculated between corresponding clusters and compared to the percent difference from the FPPS data (see Table 6).

Table 5: Model Performance of Mean Shift Clustering with FPPS

Cluster	FPPS Mean	Mean Shift Clustering Mean	Percent Error (%)	FPPS Cluster Size	Mean Shift Cluster Size (N=200)	Percent Error (%)
Cluster 1A	μ=(100, 100)	μ=(102.6, 101.2)	2.60, 1.20	100	112	12
Cluster 2A	μ=(300, 100)	μ=(296.8, 100.7)	1.07, 0.70	60	58	-3.33
Cluster 3A	μ=(200, 110)	μ=(206.1, 110.6)	3.05, 0.55	25	25	0
Cluster 4A	μ=(400, 100)	μ=(388.1, 95.7)	-3.00, -4.30	5	5	0
Cluster 1B	μ=(100, 100)	μ=(101.2, 102.7)	1.2, 2.70	110	118	7.27
Cluster 2B	μ=(300, 100)	μ=(300.0, 101.6)	0, 1.60	60	60	0
Cluster 3B	μ=(200, 100)	μ=(209.3, 104.8)	4.65, 4.80	20	22	10

MSC Cluster Mean Percent Error less than 5%; MSC Cluster Sigma Percent Error less than 12%.

Table 6: Fibroblast and TC32 Mean Shift Clustering

Cluster	Mean Shift Clustering Mean	Absolute Value of Percent Difference (x%, y%)	Mean Shift Clustering Size	Percent Difference (%)
Cluster 1A	μ=(118.11, 87.34)	0.002, 2.93	603	6.79
Cluster 2A	μ=(389.16, 68.88)	7.38, 7.51	161	19.88
Cluster 1B	μ=(118.07, 89.90)	0.002, 2.93	562	-7.29%
Cluster 2B	μ=(420.10, 64.07)	7.38, 7.51	129	-24.81%

Distribution A is Fibroblast and Distribution B is Fibroblast. The percent difference is the percent difference between the corresponding clusters from Distribution A to Distribution B. Less than 3% difference in the first cluster mean and greater than 7% difference in the second cluster mean between the distributions. Significant cluster 2 sigma difference (>20%).

Next, to validate the DRP metric, points Distribution A and Distribution B were generated with two clusters each. The parameters were as follows: Mean 1A = (100, 100), Mean 1B = (200, 100), Mean 2A = (100, 100), Mean 2B = (350, 100), with sigmas for all clusters held at 10. . The first cluster was identical for both distributions, while the second cluster differed statistically between the distributions. The DRP metric was then applied, and a Multigussian curve was fitted to determine the peak parameters. A t-test was conducted to validate that the p-value was greater than 0.05, indicating statistical indifference, and that the p-value for the second cluster was less than 0.05, confirming it was statistically different.

Acknowledgements

The author expresses gratitude to Dr. Gaudenz Danuser and Dr. Hanieh Mazloom Farsibaf at UT Southwestern Medical Center's Lyda Hill Department of Bioinformatics for their invaluable expertise and resources. Special thanks also go to Dr. Hanieh Mazloom Farsibaf for her guidance and mentorship throughout the development of this research paper.

References

1 B. Lim and G. N. Hortobagyi, *Cancer Metastasis Reviews*, 2016, **35**, 495–514.

2 M. Najafi, N. H. Goradel, B. Farhood, E. Salehi, S. Solhjoo, H. Toolee, E. Kharazinejad and K. Mortezaee, *Journal of Cellular Physiology*, 2018.

3 E. Fokas, W. G. McKenna and R. J. Muschel, *Cancer Metastasis Reviews*, 2012, **31**, 823–842.

4 A. Dominiak, B. Chełstowska, W. Olejarz and G. Nowicka, *Cancers*, 2020, **12**, 1232.

5 E. A. Sweet-Cordero and J. A. Biegel, *Science*, 2019, **363**, 1170–1175.

6 D. Saucier, X. Jiang, D. Rajendran, R. Ravishankar, E. Butler and A. Marchetto, *Scientific Reports*, 2020, **10**, 1910.

7 R. Díez-Martínez and J. Oyarzabal, *Cancer Research*, 2018, **78**, 6048–6052.

8 E. Cuppen, *Molecular Cancer Research*, 2008, **6**, 685–694.

9 Columbia University Mailman School of Public Health, *Hot spot spatial analysis*, 2023, https://www.publichealth.columbia.edu/research/population-health-methods/hot-spot-spatial-analysis#:~:text=An%20example%20is%20the%20Gettis,usually%20set%20at%2099.9%25.

10 J. Lengyel, F. Sémécurbe and S. G. Roux, *Environment and Planning B: Urban Analytics and City Science*, 2023, **51**, 3–25.

11 B. N. Boots and A. Getis, *Point pattern analysis*, WVU Research Repository, Reprint edn, 1988.

12 R. Teixeira, A. O'Connor and M. Nogal, *Structural Safety*, 2020, **82**, 101894.

13 J. Yang, E. Grunsky and Q. Cheng, *Computers & Geosciences*, 2019, **123**, 92–101.

14 B. Kranstauber, M. Smolla and K. Safi, *Methods in Ecology and Evolution*, 2017, **8**, 1559–1568.

15 T. Shibue and R. A. Weinberg, *Seminars in Cancer Biology*, 2011, **21**, 99–106.

16 A. S. Fotheringham and C. Brunsdon, *Geographical Analysis*, 1999, **31**, 340–358.

17 M. E. Davis, D. W. Tipping, R. A. Leach, E. L. Higa and M. C. Rogers, *Cancer Research*, 2016, **76**, 2871–2875.

18 S. Weglarczyk, *ITM Web of Conferences*, 2018, **23**, 00037.