Gene expression

Detecting Data Embedding Spatial Patterns and Identifying Biomarkers with BioRSP

Zeyu Yao[1,*] and Jake Y. Chen[2]

[1]AlphaMind Club, LLC, Birmingham, AL, USA 35244 and [2]Department of Biomedical Informatics and Data Science, University of Alabama at Birmingham, Birmingham, AL 35294

*To whom correspondence should be addressed.

Abstract

Motivation: High-dimensional data embedding has become an essential tool in bioinformatics, particularly for single-cell data analysis, as it aids in characterizing cells with complex gene expression patterns based on projected low-dimensional data point spatial distributions. However, traditional methods often fall short in distinguishing subtle changes within cell populations. To address this, we developed BioRSP (**Bio**logical **R**adar **S**canning **P**lots), an innovative open-source software designed to enhance the characterization of single-cell gene expression patterns by simulating radar beam scanning across defined clusters from a fixed coordinate, referred to as the "vantage point."

Results: BioRSP enables users to analyze the distribution of each gene within and across cell clusters using a quantitative measure known as the RSP plot. These plots categorize genes and spatial embedding patterns into four distinct groups based on combinations of gene coverage (high/low) and RSP values (high/low), reflecting spatial distribution regularities in the embedding space. This novel approach allows BioRSP to distinguish subtle differences between background cells and those expressing specific markers above a preset threshold, significantly improving upon traditional gene-to-gene correlation analyses. A case study utilizing a neonatal mouse heart tissue dataset from the Mouse Cell Atlas demonstrates BioRSP's capability to identify spatially distinct and functionally significant gene expression patterns.

Availability: The Python package and its comprehensive documentation are publicly available at https://github.com/cytronicoder/biorsp.

Contact: cytronicoder@gmail.com

1 Introduction

High-dimensional data embedding has emerged as a crucial tool in bioinformatics, especially for single-cell data analysis (Vandereyken et al., 2023; Wang et al., 2022). Understanding such patterns is crucial for researchers to identify cellular heterogeneity, uncover novel cell types, and decipher the spatial organization of tissues (Gupta and Kuznicki, 2020). By integrating spatial transcriptomics with advanced computational tools, researchers can visualize gene expression dynamics and interactions within their native tissue context. To address this, we focus on characterizing these spatial patterns of gene expression within cell populations in a 2D embedded space, enabling more precise analysis of gene distributions and interactions.

Traditional methods for comparing spatial gene expression patterns often rely on simple metrics such as coverage, which quantifies the density distribution of cells expressing specific biomarkers. Coverage is typically represented as a single numerical value or as its distribution across predefined bins. However, a single numerical value fails to capture the spatial arrangement or clustering of expressing cells, overlooks the heterogeneity within the cell population, and cannot distinguish between uniform and spatially restricted gene expression patterns. Newer methods involving spatial investigation address these limitations by analyzing the spatial distribution of gene expression within cell populations, offering a more nuanced understanding of gene dynamics.

We differentiate between the background pattern, referring to all cells in the population, and the foreground pattern, which includes only cells expressing specific genes above a particular threshold. We can identify and pinpoint marker genes with specific spatial expression patterns in certain subpopulations by analyzing these patterns considering the cell population shape and density distribution. Previous work, such as the Polar Gini Curve (PGC), has laid the groundwork for such analysis (Nguyen et al., 2021). PGCs are constructed by calculating the Gini coefficient between foreground and background cells at various angles, then rotating the data 360 degrees to derive a curve. This method generates both foreground and background PGCs, which can be compared to create a comprehensive gene expression landscape, distinguishing uniformly distributed genes from those with spatially restricted patterns. However, the pairwise comparison approach of PGC can overestimate inequality for genes with sparse coverage, as sparse data can significantly skew PGC results, producing high Gini coefficients that exaggerate inequality. This misrepresentation can suggest a distribution that appears unfairly weighted, which may be an artifact of sparse data rather than a true reflection of gene expression patterns.

To address these limitations, we developed BioRSP, a novel tool for spatial gene expression analysis. BioRSP extends PGC by using a radar-like scanning mechanism to evaluate gene expression from various angles around a central "vantage point," similar to how a radar scans its surroundings. This innovative approach allows for the detection of expression patterns that are not immediately apparent in traditional analyses, offering a more nuanced view of spatial heterogeneity. By systematically assessing gene expression from various angles, researchers can identify areas of dense gene expression indicative of specific cellular processes or regions where gene activity is significantly lower.

2 Methods

2.1 Algorithm description

The BioRSP algorithm starts with differentially expressed gene (DGE) data collected from RNA sequencing or other high-throughput methods, preprocessed to ensure quality and comparability. Then, a vantage point is defined within the dataset's spatial bounds to serve as the origin for polar transformation and as the reference point for the analysis. The scanning window, defined by a central angle and width in radians, determines the start and end angles of each scan. Furthermore, the resolution parameter specifies the number of bins for partitioning the angular domain in histogram calculations, while a threshold parameter defines the minimum expression level for a gene to be considered in the foreground pattern. Finally, the mode of comparison, either absolute or relative, dictates whether the foreground CDF is directly compared against the background CDF without normalization or scaled to match the total number of points in the background.

Dimensionality reduction techniques, such as t-SNE, are applied to visualize high-dimensional gene expression data in two dimensions (van der Maaten and Hinton, 2008). Clustering is performed using DBSCAN to identify distinct cell types or states within the t-SNE plot (Ester et al., 1996). Foreground and background points within the identified clusters are then translated relative to the chosen vantage point and converted into polar coordinates. Points within the scanning window are identified, and histograms for both foreground and background points are computed by partitioning the angular space into bins. CDFs for the foreground and background points are constructed by summing the histogram counts cumulatively, and depending on the mode of comparison, the CDFs are either directly compared or scaled.

A detailed description of BioRSP can be found in the Supplementary Material.

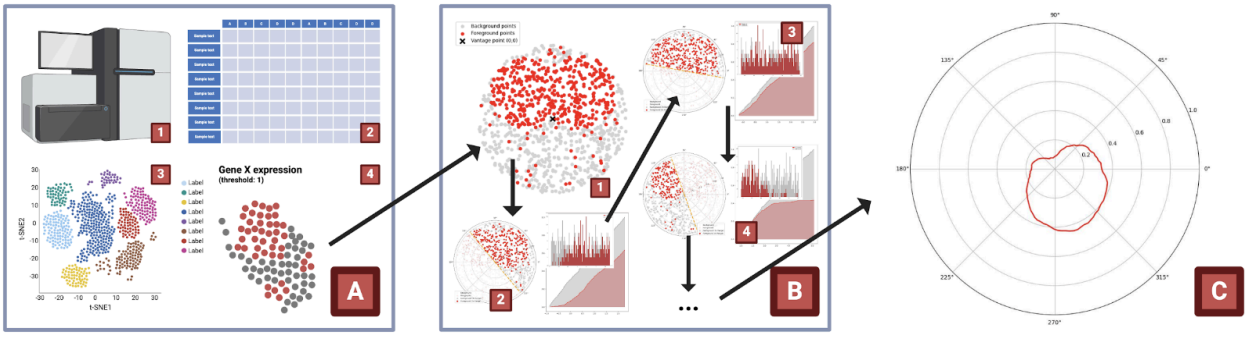

Figure 1: A summary diagram of the BioRSP workflow. (A) The workflow starts with DGE data acquisition and preprocessing, followed by dimensionality reduction using t-SNE and mapping gene expression levels onto the t-SNE plot. **(B)** BioRSP analysis involves identifying foreground and background points, transforming them into polar coordinates, generating histograms, computing cumulative distribution functions (CDFs), and calculating the Radar Scanning Plot (RSP) area to quantify spatial gene expression patterns. **(C)** The final radar plot visualizes the differences between foreground and background gene expression patterns across multiple angles.

2.2 Comparative Evaluation Simulation

To assess the effectiveness of BioRSP in identifying nuanced spatial distributions of gene coverage under varying conditions, we conducted benchmark simulations against PGC. Simulated data was generated by creating a set of points within a unit circle to represent spatial locations. Points were generated using polar coordinates (r_s, θ_s), where $r_s = \sqrt{U(0,1)}$ is the radius obtained from a uniform distribution scaled to the circle's radius and θ_s is uniformly distributed between 0 and 2π radians. Cartesian coordinates (x, y) were subsequently calculated as $x = r_s \cos(\theta_s)$ and $y = r_s \sin(\theta_s)$. We varied the foreground coverage in increments of 10%, 25%, 50%, 75%, and 90%. Bias in the foreground was introduced by varying the center of bias across five coordinates: (0, 0), (0.25, 0.25), (-0.25, -0.25), (0.25, -0.25), and (-0.25, 0.25). Points were either randomly selected for unbiased simulations or selected based on their distance from a specified center (c_x, c_y) for biased simulations, where Euclidean distances were computed as $\sqrt{(x - c_x)^2 + (y - c_y)^2}$. Points within a threshold distance, determined by the coverage parameter, were classified as foreground.

2.3 Experimental Validation

To further validate BioRSP, we applied it to a real dataset of neonatal mouse heart tissue from the Mouse Cell Atlas 3.0 to identify significant cluster markers (Wang, Zhang, et al., 2022). The dataset included the expression of 5,506 genes in 10,000 cells after preprocessing. Using DBSCAN on the generated t-SNE embedding, we identified three distinct cell clusters. We focused on all three clusters to comprehensively evaluate BioRSP's capabilities.

For each cluster, we identified significant gene pairs based on their spatial expression patterns. We extracted the genes with significant differences in spatial distribution from the bioRSP results, focusing on the pairs with p-values less than 0.05 after conducting Pearson correlation. From the significant gene pairs identified, a custom network was constructed, where nodes represented genes and edges represented significant spatial relationships.

The baseline PPI network was constructed using data from the STRING database, and a subgraph of it was extracted using the genes present in the custom network (Szklarczyk et al., 2022). This subgraph included all known interactions among the significant genes identified by bioRSP, allowing us to analyze these genes within the context of established protein interactions.

3 Results

The benchmark simulations compared BioRSP with the Polar Gini Curve (PGC), revealing BioRSP's superior ability to accurately reflect both uniform and biased distributions of gene expression (Fig. 2). In scenarios with uniform distribution, Radar Scanning Plots (RSPs) displayed circular shapes, accurately reflecting the even spread of gene expression across the spatial domain. In contrast, for biased distributions, the radii in RSPs complemented the actual spatial distribution of gene expression: Smaller or near-zero radii corresponded to high concentrations of gene expression, while larger radii indicated areas with less or no gene expression.

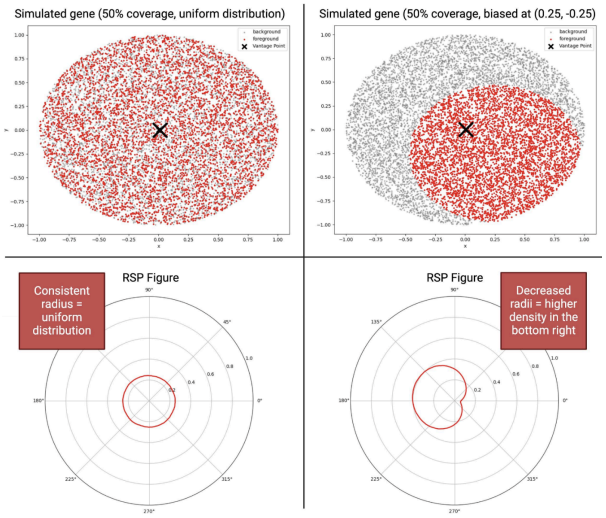

Figure 2: Simulated cell populations at 50% gene coverage with varying bias. (Top row) Distribution scatter plots (even, biased distribution towards the bottom right). **(Bottom row)** Corresponding RSP figures with documented parameters.

We also explored the effect of changing the vantage point (Fig. 3). The default vantage point was set at (0, 0), and an experimental vantage point was set at (-1, 1). Analyzing the data from these different vantage points revealed that what appears as a uniform distribution from one angle could actually be a highly concentrated cluster of gene expression when viewed from another. This finding highlights how different vantage points can reveal hidden spatial heterogeneity.

Furthermore, the scanning window width can be tuned to extract different levels of detail from the spatial expression data (Fig. 4). Smaller scanning ranges revealed more granular details in cellular expression patterns, with tighter curves illustrating high resolution, subtle variations, and localized clusters of gene expression that may indicate specific cellular states or microenvironments. Conversely, larger scanning windows smoothed out fine details to highlight broader patterns of expression across the cluster.

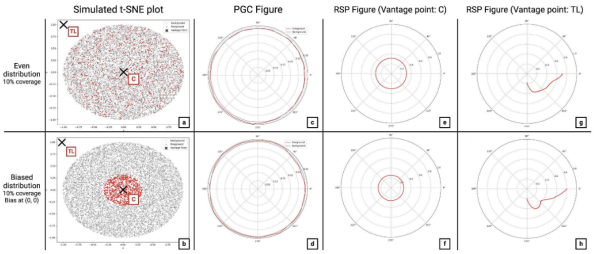

Figure 3: Vantage point variability and its impact on RSP analysis. (Top row) Analysis for a simulated gene with 10% coverage and uniform distribution. **(Bottom row)** Biased distribution of simulated gene at (0,0). **(Leftmost panel)** t-SNE plot. **(Left panel)** Corresponding PGC figure. **(Right panels)** RSP figures from the center and top-left vantage points.

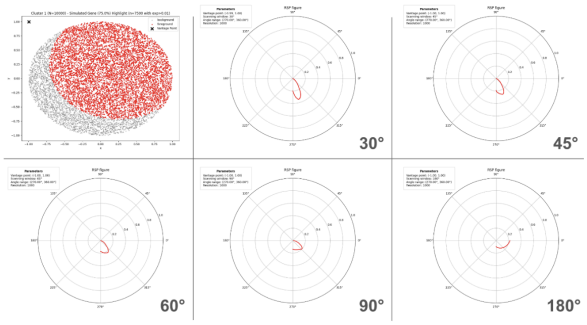

Figure 4: Exploring different scanning window sizes. Displayed is the spatial distribution of a simulated gene with 75% expression in a cellular cluster, along with RSP figures corresponding to various scanning window sizes (30°, 45°, 60°, 90°, and 180°).

3.4 Significant Gene Comparison with PPI Network

The custom network for Cluster 2 demonstrated higher specificity, with more pronounced functional modules compared to the baseline PPI subgraph (Fig. 5). For example, genes like *Sdpr* and *Tspan7* formed central hubs in the custom network for one cluster and showed numerous significant interactions with other genes involved in extracellular matrix and cellular structural organization. The custom network's focus on spatially distinct relationships also suggests better context-specific biological insights.

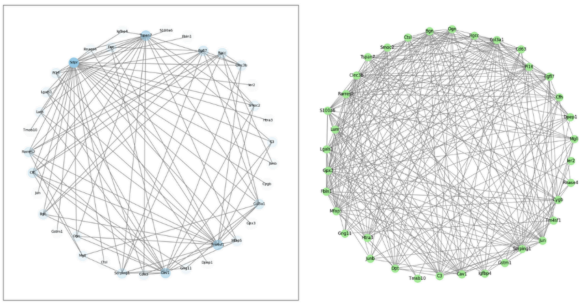

Figure 5: Comparative network analysis for cluster 2. (Left panel) Custom network constructed based on bioRSP results. **(Right panel)** Corresponding subgraph extracted from the baseline PPI network, derived from the STRING database, capturing known protein-protein interactions.

Raw data and significant gene pairs used to construct the networks, along with visualizations of the networks for each cluster and the Pearson coefficient matrices, can be found in the Supplementary Material.

4 Conclusion

The BioRSP method introduces a novel approach for analyzing spatial gene expression patterns, addressing the limitations of traditional methods like PGC. Traditional analyses often overlook spatial and complex heterogeneity within datasets. In contrast, BioRSP's radar-like scanning mechanism provides deeper insights into potential biomarkers by evaluating data from multiple angles. For instance, BioRSP could identify tumor microenvironments in oncology by highlighting regions of dense gene expression related to cancer proliferation and metastasis. Additionally, incorporating machine learning algorithms could automate the identification of significant spatial expression patterns and predict their biological functions.

References

Ester,M. et al. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In, Second International Conference on Knowledge Discovery and Data Mining. AAAI Press, Portland, Oregon, pp. 226–231.

Gupta,R.K. and Kuznicki,J. (2020) Biological and Medical Importance of Cellular Heterogeneity Deciphered by Single-Cell RNA Sequencing. Cells, 9, 1751.

Nguyen,T.M. et al. (2021) Polar Gini Curve: A Technique to Discover Gene Expression Spatial Patterns from Single-cell RNA-seq Data. Genomics, Proteomics & Bioinformatics, 19, 493–503.

Szklarczyk,D. et al. (2022) The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Research, 51, D638–D646.

van der Maaten,L. and Hinton,G. (2008) Visualizing Data using t-SNE. Journal of Machine Learning Research, 9, 2579–2605.

Vandereyken,K. et al. (2023) Methods and applications for single-cell and spatial multi-omics. Nature Reviews Genetics, 1–22.

Wang,R. et al. (2022) Construction of a cross-species cell landscape at single-cell level. Nucleic acids research, 51, 501–516.

Wang,R. et al. (2022) Integration of computational analysis and spatial transcriptomics in single-cell study. Genomics, Proteomics & Bioinformatics.

Author contributions

JC conceived the idea and the technology framework and designed the experiments and evaluation framework. He also provided guidance and feedback throughout the project. ZY implemented the idea, wrote the software codes, processed the data sets, and performed data analysis.

Competing interest statement

The authors declare that they have no competing interests related to this work.