scatterbar: an R package for visualizing proportional data across spatially resolved coordinates

Dee Velazquez[1,2] and Jean Fan[1,2]

[1]Center for Computational Biology, Whiting School of Engineering, Johns Hopkins University, Baltimore, MD 21211, USA

[2]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218, USA

Correspondence should be addressed to:

Jean Fan (jeanfan@jhu.edu)

Key words: data visualization, spatial transcriptomics, spatial analysis

Abstract

Summary

Displaying proportional data across many spatially resolved coordinates is a challenging but important data visualization task, particularly for spatially resolved transcriptomics data. Scatter pie plots are one type of commonly used data visualization for such data but present perceptual challenges that may lead to difficulties in interpretation. Increasing the visual saliency of such data visualizations can help viewers more accurately identify proportional trends and compare proportional differences across spatial locations. We developed scatterbar, an open-source R package that extends ggplot2, to visualize proportional data across many spatially resolved coordinates using scatter stacked bar plots. We apply scatterbar to visualize deconvolved cell-type proportions from a spatial transcriptomics dataset of the mouse olfactory bulb to demonstrate how scatter stacked bar plots can enhance the distinguishability of proportional distributions compared to scatter pie plots.

Availability and implementation

Scatterbar is available at https://github.com/JEFworks-Lab/scatterbar with additional documentation and tutorials at https://jef.works/scatterbar/.

1 Introduction

Effective data visualization plays a pivotal role in uncovering insights and communicating complex information clearly[1]. This is particularly important when dealing with proportional data in addition to spatial coordinates, as it combines quantitative values with their spatial context, providing a more comprehensive understanding of the underlying phenomena present and noting certain areas of interest. Such visualization of proportional data in addition to spatial coordinates is particularly important for spatial transcriptomic (ST). ST enables high-throughput profiling of gene expression within tissue sections, often at multi-cellular pixel resolution[2,3,4]. Such data demand appropriate visualizations to accurately represent the presence of multiple cell types and their respective proportions in each pixel while preserving their spatial coordinates. Detecting cell-type-specific spatial variation such as proportional changes across spatial locations is important for understanding tissue organization and disease mechanisms. Therefore, there is a need for more effective visualization techniques that enhance the saliency of proportional changes and make it easier to interpret such complex spatial datasets.

Currently, a widely used method for this visualization is the scatterplot of pie charts, commonly referred to as “scatter pie plots” as implemented through the scatterpie R package[5]. Scatter pie plots offer a way of displaying proportions at specific spatial locations by combining the spatial plane of scatterplots with the proportionate data representation of pie charts. However, the limitations of using pie charts to accurately visualize quantitative proportional information have been previously well characterized[6]. Pie charts encode quantitative data using angles, resulting

in decreased accuracy in data interpretation compared to if the quantitative data was encoded using lengths. Previous studies have shown that accurately comparing the sizes of the slices in a pie chart, especially when the differences are subtle, is more challenging than comparing the lengths of bars, for example, in a bar chart. As such, the use of scatter pie plots for ST data visualization may make it challenging to accurately interpret and compare cell-type proportions to identify potentially interesting changes across spatial locations.

2 Materials and methods

Here, we present an R package called 'scatterbar,' which provides an alternative visualization for displaying proportional data across many spatially resolved coordinates. In contrast to scatterpie, scatterbar uses stacked bar charts instead of pie charts. Given a set of (x,y) coordinates and matrix of associated proportional data, the package creates a stacked bar chart, where bars are stacked based on the proportions of different categories centered at each (x, y) location.

The package's core functionality is in the function 'create_scatterbar', which first verifies that the two input data frames are correctly formatted. The input data frames comprise of a data frame containing the proportions of different categories for each location, where each row represents a location in a 2-D plane and each column represents a category, and a data frame containing the positional information for each location (specified in the row names). It then

reshapes the proportion data from a wide format to a long format using the tidyr package[7]. The positional data is prepared by renaming the columns to x and y and ensuring that the location identifiers match between the proportion data and the positional data.

Subsequently, the function filters the locations to include only those that exist in both data frames, merges the reshaped proportion data with the positional data, and calculates the cumulative proportions for each location using the dplyr package[8] . This step ensures that the heights of the bars within a spot add up to 1, maintaining the relative proportions of each category.

To modulate the size of the bar plots, scaling factors are used. Scaling factors are multipliers applied to the axes to adjust the size of the bars relative to the plot's coordinate system. They are crucial as they enable the visualization to fit within the specified plot area and thereby make the proportions more interpretable. If the scaling factors for the x and y axes are not provided, the function automatically calculates optimal scaling factors based on the range of the coordinates and the number of locations provided. Padding for the x and y axes can also be adjusted as needed. Padding refers to the additional space added around the plot elements to prevent overlap with the plot's boundaries or other elements. Adjusting the padding ensures that the bars are clearly visible and separated from the edges of the plot.

The final scatter bar plot is generated using the grammar of graphics framework as implemented **in** ggplot2[9], with options for customizing the colors of the bars, the title of the

plot, and the legend. The function returns a ggplot object representing the scatter bar plot, which can be further customized and rendered using standard ggplot2 extensions such as ggthemes for additional aesthetics.

3 Usage scenario

To illustrate the potential utility of the scatterbar package, we created both the scatter pie and scatter bar plots to visualize a multi-cellular pixel resolution ST dataset of the mouse olfactory bulb[2]. Briefly, for this dataset, gene expression measurements were profiled at 55µm resolution pixels across the tissue. Using these gene expression measurements, cell-type proportions per pixel were estimated using the reference-free deconvolution approach STdeconvolve[10]. The input to scatterpie and scatterbar thus contained the deconvolved cell-type proportions at each pixel and the corresponding centroid spatial coordinates of each pixel. As such, we aimed to use data visualization to visualize and compare the deconvolved cell-type proportions across spatial locations in the tissue.

While both scatter pie and scatter bar plots allow for the visualization of cell-type proportions across spatial locations, it can be challenging to discern subtle differences in cell-type proportions between locations based on the scatterpie plot due to the visual similarity of angles in pie charts. In particular, we highlight two spatial locations where the green cell-types are equal in proportion, but the red cell-type is not. In visually assessing the green and red areas

denoting the proportion of the green and red cell-types respectively in the pie charts in the scatter pie plot, we note that the slice areas appear identical suggesting that the proportion of the green and red cell-type is equal across these two spatial locations in the tissue (Fig 1A). Alternatively, in visually assessing the green and red bars denoting the proportion of the green and red cell-types respectively in the bar charts in the scatter bar plot, we can more readily discern that the proportion of the red cell type is slightly different at the two locations (Fig 1B).

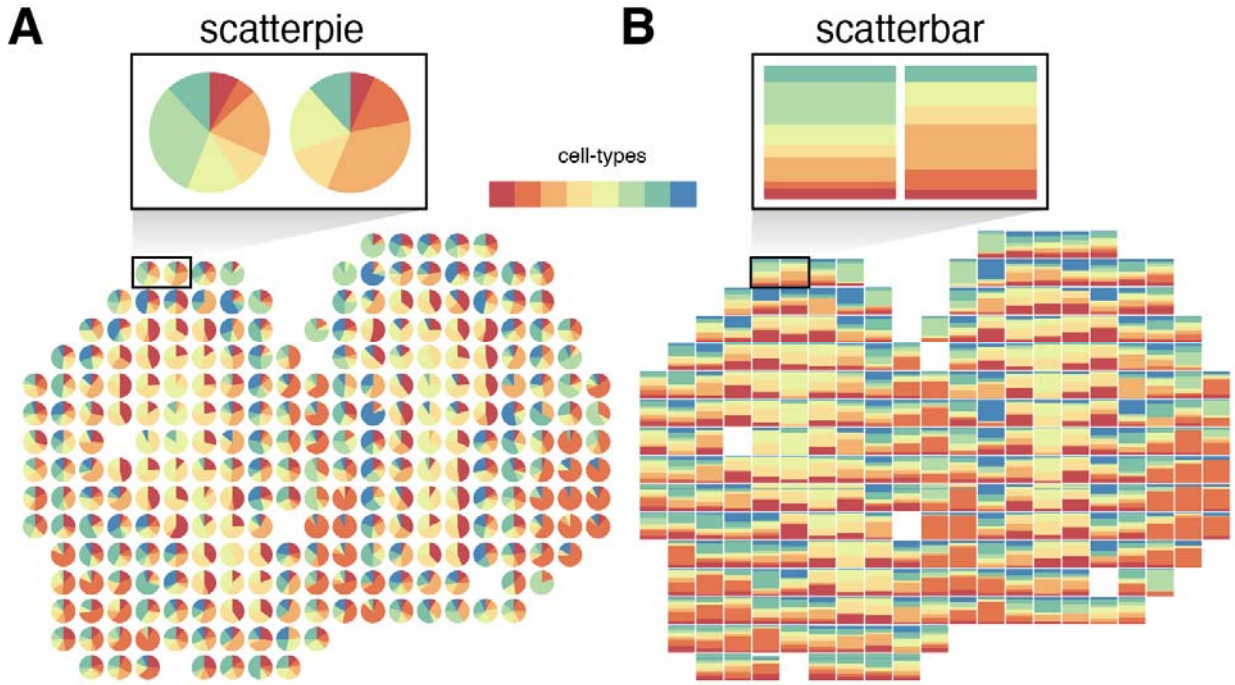

Figure 1. Comparison of scatter pie and scatter bar plots for deconvolved cell-type proportions in spatial transcriptomics data of the mouse olfactory bulb. (A) Scatterpie plot showing the proportions of eight cell types across spatial coordinates. Each pie chart represents the composition of cell types at that specific spatial location in the tissue, with the size of the slice of each pie segment corresponding to the proportion of each cell type. (B) Scatterbar plot showing the same data, where each spot is represented by a stacked bar chart. Each bar chart

represents the composition of cell types at that specific spatial location in the tissue, with the height of each bar segment corresponding to the proportion of each cell type. The x-axis and y-axis represent the spatial coordinates, and the colors correspond to the different cell types. Inset highlights two spatial locations.

By using bar charts instead of pie charts, scatterbar enhances the saliency of cell-type proportion differences between spots, making it easier to detect subtle variations in cell-type distribution across spatial locations in the tissue. Our observations align with the insights reported by researchers who have performed perceptual studies on data visualizations, reinforcing the effectiveness of the scatterbar approach in visualizing spatial transcriptomics data[1,6].

4 Conclusion

We presented scatterbar as an R package implementing scatter bar plots for visualizing proportional data across spatially resolved coordinates. Compared to more commonly used scatter pie plots, scatter bar plots can enhance the saliency and distinguishability of proportional distributions, making it potentially easier to interpret and compare data across different spatial locations. By extending ggplot2, scatterbar allows users to readily customize and tailor the visualization to specific datasets and research needs. Although we have focused on demonstrating the utility of scatterbar in the visualization of deconvolved cell-types in ST data, accurately visualizing changes in proportional data is crucial for a variety of applications,

including but not limited to geographic information systems (GIS), phylogeography, in addition to biological research. As such, we anticipate that scatterbar will be a relevant tool for a wide range of applications.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 2047611. We thank Mayling Chen, Caleb Hallinan, Srujan Singh, and Kamil Slowikowski for testing and feedback of the software package.

References

1. Wong, B. (2010). Design of data figures. Nature Methods.
2. Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., Mollbrink, A., Linnarsson, S., Codeluppi, S., Borg, Å., Pontén, F., Costea, P. I., Sahlén, P., Mulder, J., Bergmann, O., … Frisén, J. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science, 353(6294), 78–82. https://doi.org/10.1126/SCIENCE.AAF2403
3. Rodriques, S. G., Stickels, R. R., Goeva, A., Martin, C. A., Murray, E., Vanderburg, C. R., Welch, J., Chen, L. M., Chen, F., & Macosko, E. Z. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science, 363(6434), 1463–1467. https://doi.org/10.1126/SCIENCE.AAW1219
4. Liu, Y., Yang, M., Deng, Y., Su, G., Enninful, A., Guo, C. C., Tebaldi, T., Zhang, D., Kim, D., Bai, Z., Norris, E., Pan, A., Li, J., Xiao, Y., Halene, S., & Fan, R. (2020). High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue. Cell, 183(6), 1665-1681.e18. https://doi.org/10.1016/J.CELL.2020.10.026
5. Yu G, Xu S (2024). scatterpie: Scatter Pie Plot. R package version 0.2.3, https://github.com/YuLab-SMU/scatterpie/tree/master, https://cran.r-project.org/web/packages/scatterpie/vignettes/scatterpie.html
6. Mackinlay, Jock. (1986). Automating the Design of Graphical Presentations of Relational Information. ACM Trans. Graph.. 5. 110-141. 10.1145/22949.22950.

7. Wickham H, Vaughan D, Girlich M (2024). tidyr: Tidy Messy Data. R package version 1.3.1, https://github.com/tidyverse/tidyr, https://tidyr.tidyverse.org.
8. Wickham H, François R, Henry L, Müller K, Vaughan D (2023). dplyr: A Grammar of Data Manipulation. R package version 1.1.4, https://github.com/tidyverse/dplyr, https://dplyr.tidyverse.org.
9. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
10. Miller, B. F., Huang, F., Atta, L., Sahoo, A., & Fan, J. (n.d.). Reference-free cell type deconvolution of multi-cellular pixel-resolution spatially resolved transcriptomics data. https://doi.org/10.1038/s41467-022-30033-z