scatterHatch: an R/Bioconductor package for color blind accessible visualization of single-cell data

Tejas Guha[1], Elana Fertig[2-6], Atul Deshpande[2-4]

[1]Department of Electrical and Computer Engineering, A. James Clark School of Engineering, The University of Maryland, College Park, 20742, MD, USA

[2] Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA

[3]Convergence Institute, Johns Hopkins University, Baltimore, MD, USA

[4]Bloomberg-Kimmel Immunotherapy Institute, Johns Hopkins University School of Medicine, Baltimore, MD, USA

[5]Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD, USA

[6]Department of Biomedical Engineering, Johns Hopkins University School of Medicine, Baltimore, MD, USA

Abstract

Summary: Color is often used as a primary differentiating factor in visualization of single-cell and multi-omics analyses. However, color-based visualizations are extremely limiting and require additional considerations to account for the wide range of color perceptions in the population. The *scatterHatch* package provides software for accessible single-cell visualizations that use patterns in conjunction with colors to amplify the distinction between different cell types, states, and groups.

Availability: *scatterHatch* is available on Github at https://github.com/FertigLab/scatterHatch.

Contact: adeshpande@jhu.edu.

Supplementary information: Supplementary figures are available in the attached document.

Introduction

Data visualization is a central challenge in the analysis of single-cell datasets, which require methods for distinguishing among a variety of factors associated with cell types and experimental conditions. Often, these analyses rely on plotting distinct cell groups within embeddings or reduced dimension scatter plots. Despite the high complexity of single-cell visualizations, color often remains the sole visual indicator of cell groups. However, with 8% of men and 0.5% of women (Wong, 2011) affected by color vision deficiency (CVD), using color as the sole differentiating feature in visualizations limits the interpretability of data and results. In fact, the interpretability of visualizations degrades even for readers with normal vision as the number of colors increase (see Supplementary Fig. S1). In the specific context of single-cell data, we commonly use point-based visualizations where each cell is represented by a point in two-dimensional space whose coordinates correspond to either a low-dimensional embedding or the physical location of the cell in tissues. The availability of R packages to simulate different CVD's *(Ou, 2021)*, or to generate visualizations with colorblind friendly color palettes (Bunis et al., 2020) is a step in the right direction. However, these colorblind-friendly palettes may not be equally accessible to all forms of CVD, underlining the need for additional visual cues. We can use different point shapes to differentiate cell groups in sparse scatter plots, but individual point shapes cannot be deciphered in dense clusters. On the other hand, packages that implement hatching patterns over large polygonal areas are not suitable for sparse point distributions. We present *scatterHatch*, an R package for generating easily interpretable single-cell point visualizations by using a combination of patterns and colors as visual cues.

Package Description

scatterHatch adds patterns to points on a scatter plot to provide an additional visual cue to differentiate factors (see Fig. 1). *scatterHatch* is developed based on the *ggplot2* framework to visualize reduced-dimension scatter plots (e.g., PCA, UMAP, tSNE, etc.) and the spatial layout of cells in a tissue sample. It can handle hybrid point distributions with a mixture of irregular sparse and dense point distributions by using a simple pattern library which is applicable for individual points as well as for hatching patterns over large polygonal areas. For each factor, we classify cells as solitary points or group them into large and small point clusters based on the nearest neighbor distances from their corresponding point positions in the visualization. For instance, the nearest neighbor distances in a UMAP plot are calculated using the distances between the cells in the low-dimensional UMAP space. We used the *spatstat (Baddeley et al., 2015)* package for finding the nearest neighbor distances. We plot a coarse grid pattern over the large and small clusters in the dense regions. In the sparse regions consisting of cells classified as solitary points, we individually plot the pattern on each point.

scatterHatch has six standard patterns - horizontal, vertical, right diagonal, left diagonal, checkers, and crisscross - in addition to supporting a "blank" pattern where no lines are drawn for a group. The default color palette has 40 CVD friendly colors from the *dittoSeq* package. Even for a limited color palette, *scatterHatch* can enhance the interpretability of a point visualization by assigning multiple patterns for the same color. Thus, *scatterHatch* can support up to 280 individual groups through combinations of the default patterns and colors. With 82 groups (see Supplementary Fig. S2), *scatterHatch* allows clear distinction between groups with visually similar colors when compared to plots without patterns (see Supplementary Fig. S3) for varying color perceptions. On the other hand, *scatterHatch* can display scatter plots with as few groups as four in a visually pleasing and distinct way (see Supplementary Fig. S4). Users can increase the number of patterns beyond the default by customizing patterns using different line types (e.g., dashed, dotted) and line colors. In addition, users can define a new pattern composed of multiple lines by providing a set of corresponding line angles (see Supplementary Fig. S5).

Application Example

To illustrate the application of *scatterHatch* to enhance CVD efficiency, we use single-cell data from a resection specimen (Lin et al. 2018) of Pancreatic Ductal Carcinoma (PDAC) and adjacent normal tissues. We selected 10,000 cells at random from this dataset and calculated a reduced dimension representation using UMAP. We classified these cells into four groups using K-means clustering and generated a UMAP plot using *scatterHatch*. We simulated the distinguishability between the cell groups perceived by CVD individuals using the "cvdPlot" function from the *colorBlindness* package (Figure 1, script available at https://github.com/FertigLab/scatterHatch-paper). This example demonstrates the use of texture to facilitate distinction between cell groups, irrespective of the associated colors. As demonstrated by the simulated monochromatic figure, adding patterns to points enables reliable distinction between cell groups for sparse and dense regions of the plot.

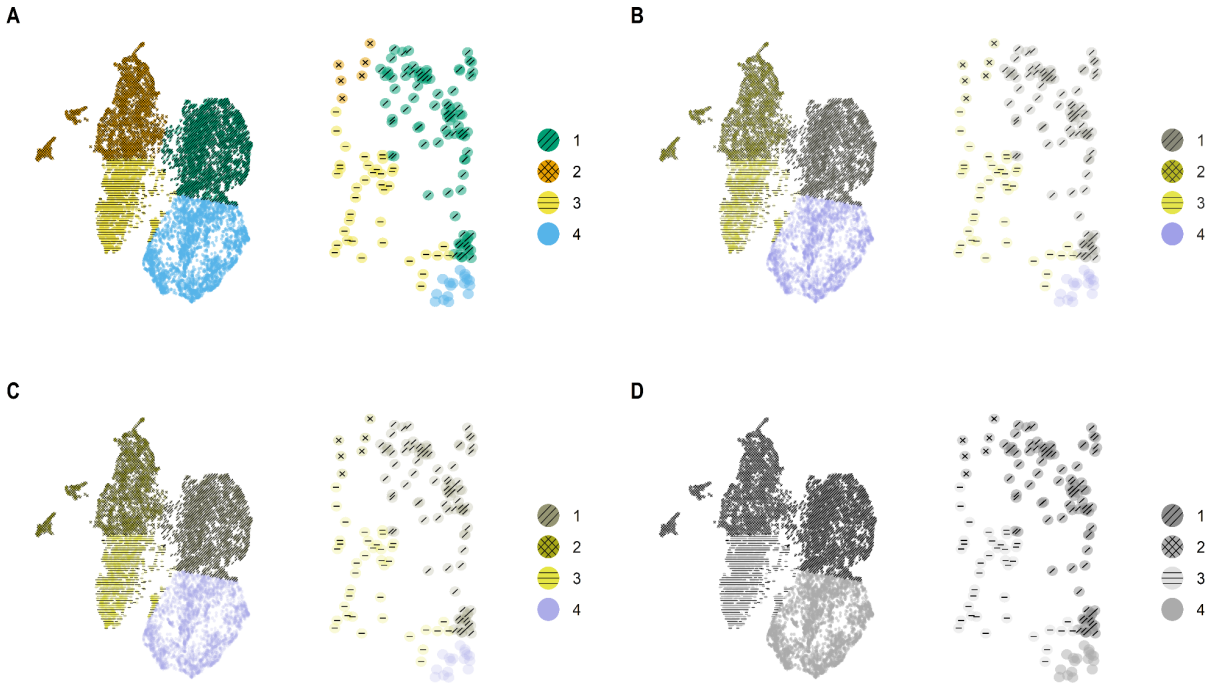


Fig. 1. Simulated perception of a UMAP plot generated using scatterHatch by individuals with (A) normal vision, (B) deuteranomaly, (C) tritanomaly, (D) monochromacy, with a sparse region magnified to show patterns assigned to individual cells.

Conclusion

scatterHatch is an R package for generating colorblind friendly point visualizations of embeddings for single-cell and spatial datasets. We demonstrate how *scatterHatch* facilitates distinction between groups of cells in low dimensional scatter plots (e.g., PCA, UMAP, TSNE) as well as visualizations of cellular distribution within the tissue. In the future, *scatterHatch* will be expanded by broader compatibility with data structures from packages commonly used in single cell and computational biology workflows like the Single Cell Experiment Object from *Seurat*. Additionally, *scatterHatch* will be improved to work better with visualizations with overlapping cell groups seen in many tSNE and UMAP plots. In conclusion, *scatterHatch* introduces a single function that allows a user to build scatter plots with two visual cues - color and pattern - that are as aesthetically pleasing as accessible to an audience of varying color perception.

Funding

This work was supported by National Institutes of Health grants U01CA253403, U01CA212007, and P01CA247886.

Conflict of Interest: The authors have no conflicts of interest to declare.

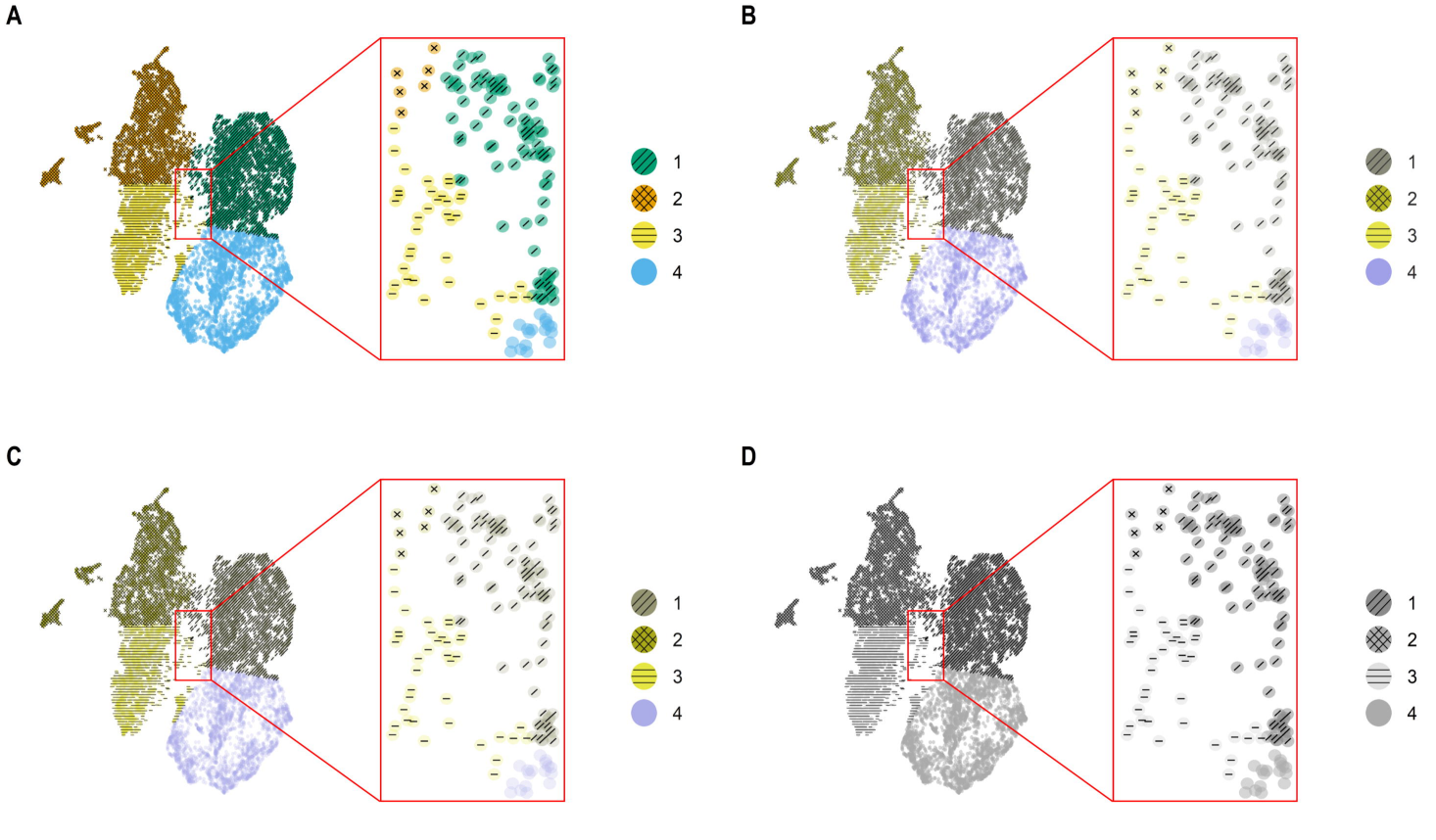
References

Baddeley, A., Rubak, E., & Turner, R. (2015). *Spatial Point Patterns: Methodology and Applications with R*. CRC Press.

Bunis, D. G., Andrews, J., Fragiadakis, G. K., Burt, T. D., & Sirota, M. (2020). dittoSeq: Universal User-Friendly Single-Cell and Bulk RNA Sequencing Visualization Toolkit. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btaa1011

Ou, J. (2021, April 16). *colorBlindness Guide*. https://cran.r-project.org/web/packages/colorBlindness/vignettes/colorBlindness.html

Wong, B. (2011). Color blindness. *Nature Methods*, *8*(6), 441.


A
1
2
3
4
B
1
2
3
4
C
1
2
3
4
D
1
2
3
4