FlexDotPlot: a universal and modular dot plot visualization tool for complex multifaceted data

Simon Leonard[1,2], Antoine Rolland[1], Karin Tarte[2,3], Frédéric Chalmel[1,*] and Aurélie Lardenois[1]

[1] Univ Rennes, Inserm, EHESP, Irset (Institut de recherche en santé, environnement et travail) - UMR_S 1085, F-35000 Rennes, France.
[2] UMR 1236, University of Rennes, INSERM, Etablissement Français du Sang Bretagne, Rennes, France
[3] SITI Laboratory, Etablissement Français du Sang Bretagne, CHU Rennes, Rennes, France
[*] Correspondence: frederic.chalmel@inserm.fr

Abstract

Motivation: Dot plots are heatmap-like charts that provide a compact way to simultaneously display two quantitative information by means of dots of different sizes and colours. Despite the popularity of this visualization method, particularly in single-cell RNA-seq studies, existing tools used to make dot plots are limited in terms of functionality and usability.
Results: We developed FlexDotPlot, an R package for generating dot plots from any type of multifaceted data, including single-cell RNA-seq data. FlexDotPlot provides a universal and easy-to-use solution with a high versatility. An interactive R Shiny application is also available in the FlexDotPlot package allowing non-R users to easily generate dot plots with several tunable parameters.
Availability and implementation: Source code and detailed manual are available at **https://github.com/Simon-Leonard/FlexDotPlot**. The Shiny app is available as a stand-alone application within the package.

Introduction

Visualization of complex multifaceted data is essential for the biological mining of the vast amount of information generated by high-throughput technologies. One of the most popular plotting technique in genomics is the heatmap. It is used to display a single quantitative information from a two-dimensional table (Wilkinson and Friendly, 2009). For instance, in transcriptomics, rows and columns usually represent genes and samples, respectively, and boxes are colour-coded according to expression signals. Heatmaps however appear intrinsically limited to provide comprehensive representation of the diversity of information now available from single-cell technologies. Accordingly, alternative methods such as dot plots are increasingly used.
A dot plot is a modified heatmap where each box in the grid is replaced by a dot. In addition to quantitative values that are displayed via a colour gradient, the dot size is also used to represent another quantitative information, e.g. the fraction of cells expressing a given gene within a single-cell RNA-sequencing (scRNA-seq) dataset (Lukassen et al., 2018; Habib et al., 2017; Ordovas-Montanes et al., 2018; Wu et al., 2018).

In recent years, multiple tools have been developed to make dot plots like the rain plot method (Henglin et al., 2019) or the feature-expression heat map method (Benno Haarman et al., 2015; Wei and Simko, 2017). Several programs dedicated to scRNA-seq analysis (Seurat, scClustViz or cellphonedb) also provide a dot plot function (Innes and Bader, 2019; Stuart et al., 2019; Efremova et al., 2019). A dot plot generator is also available in ProHits-viz, a web-tool dedicated to protein-protein interaction analysis (Knight et al., 2017).
While the increasing number of dedicated tools embedding dot plot representation clearly illustrates the relevance of these visualization methods in current data analysis, dot plots remain difficult to generate partly due to the lack of universality. In particular, most of the previously mentioned packages use very different input data formats: ProHits-viz web-tool requires a file with quantitative information on bait-prey interactions (typically SAINT or CRAPome results (Choi et al., 2011; Mellacheruvu et al., 2013)); cellphonedb requires an output of the cellphonedb pipeline; Seurat and scClustViz are restricted to scRNA-seq data and have their own single-cell object format; corrplot (used in feature-expression heat map) and rain plot are respectively restricted to correlation and association matrixes. In addition, all of the existing methods can only display two quantitative information per dot.

In an effort to address these issues and provide the scientific community with a simple and universal tool for dot plot visualization, we developed the FlexDotPlot package.

Implementation

FlexDotPlot is implemented in R and takes advantage of several public R packages for data visualization (ggplot2, ggforce, ggdendro, grDevices, gtable, grid, gridExtra, shiny, shinydashboard, shinyWidgets, colourpicker), manipulation (reshape2) and analysis (FactoMineR) (Wickham et al., 2019; Pedersen, 2019; de Vries and Ripley, 2016; Wickham and Pedersen, 2019; Auguie, 2017; Wickham, 2017; Husson et al., 2019; Chang et al., 2019; Chang and Borges Ribeiro, 2018; Perrier et al., 2019; Attali, 2017). It requires a simple matrix as input: the first two columns contain the two factors to spread along the x and y axes (e.g. genes and cell populations for scRNA-seq datasets), followed by the corresponding quantitative (continuous) and/or qualitative (discrete) data to be displayed. From this input, dot plots can be produced with a single command line or interactively with the Shiny application. FlexDotPlot is very useful for scRNA-seq data but it can be used to describe any other large and complex dataset.

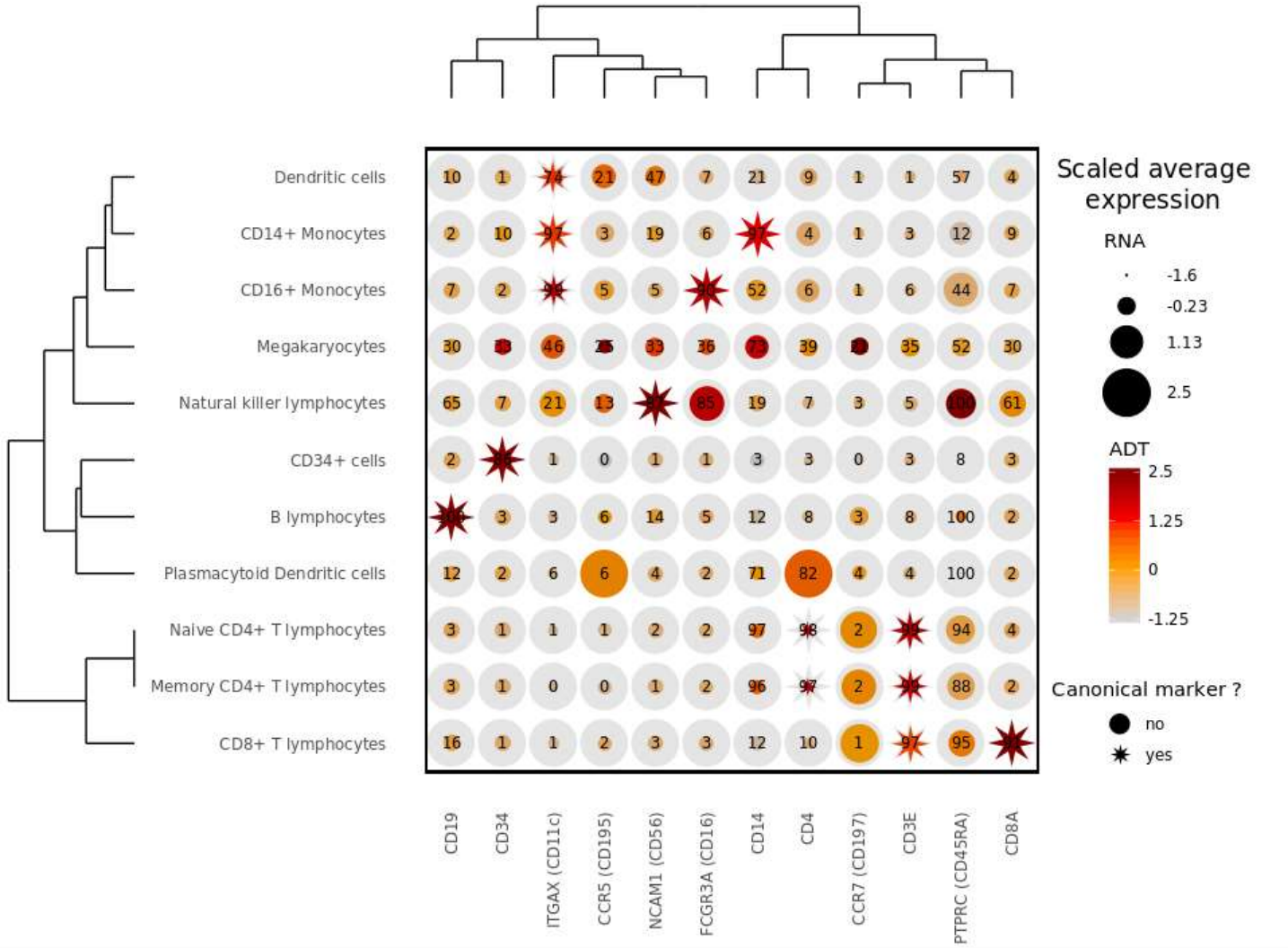

Figure 1 : FlexDotPlot representation applied to the 8k human CBMC dataset. Each dot represents multiple features for one gene (vertical axis) in a given cell cluster (horizontal axis). The size and the colour of each dot represent the average scaled expression of a given marker at the RNA level and at the Antibody Derived Tags (ADT) level, respectively. Numerical values on each dot represent the percentage of cells for which the marker is detected according to ADT. The shape of the dot is used to highlight canonical marker of specific cell types. The dataset was downloaded from the Seurat website and was processed as described *(https://satijalab.org/seurat/v3.1/multimodal_vignette.html)*.

Features highlights

FlexDotPlot consists of a single function to generate dot plots with several easy-to-tune parameters allowing users to specify which and how information has to be represented. In addition to the traditional size and colour parameters, users can display two additional information by adding some text or by using dot shapes (Figure 1). Size parameter has to be controlled by a continuous variable while others can be continuous or discrete. Several minor parameters are implemented in the function to easily custom the plot legends, shapes, colours or labels. Moreover, a ggplot2 object can be returned so that users can add custom layers if needed.

FlexDotPlot also provides parameters to enhance the resulting figure with dendrograms. Depending on the variable types, a principal component analysis (continuous variables), a multiple correspondence analysis (discrete variables) or a factor analysis for mixed data (continuous and discrete variables) can be performed prior to clustering analysis, the latter being also tunable. To improve flexibility, all of the previously described parameters are optional.

An interactive Shiny application also allows users to run FlexDotPlot without command line, directly in a browser. From the import tab, an example dataset can be loaded. From the visualization tab, the user can select x/y grid axis and variables to control parameters (size, colour, shape and text). Once the plot is displayed, additional functions and graphical parameters remain adjustable. The resulting plot can be directly saved in multiple formats while the corresponding R code is printed.

Conclusion

FlexDotPlot is a user-friendly and flexible tool for dot plot generation allowing users to independently visualize up to four parameters. By making it compatible with a simple matrix input and by offering a Shiny interface, we incentivize the use of dot plots and enable users without programming skills to visualize any kind of data. Finally, the integration

of FlexDotPlot with ggplot2 allows developers to customize even further their output.

Acknowledgments

We thank Paul Rivaud for his critical reading. FlexDotPlot is supported, built, and maintained by the Research Institute for Environmental and Occupational Health (IRSET) and the GenOuest BioInformatics core facility. We thank the Institut national de la santé et de la recherche médicale (Inserm) and the Université de Rennes 1 for supporting this work. S.L is supported by the chair "Cancer & Innovation" of Rennes 1 Foundation https://fondation.univ-rennes1.fr/.

Fundings

This work was supported by the Swiss National Science Foundation [SNF n° CRS115_171007], the French National Institute of Health and Medical Research (Inserm) [HuDeCA project to F.C.], the European Commission [HuGoDeCA project n° 874741 to A.D.R.], the University of Rennes 1, the French School of Public Health (EHESP), and the chair "Cancer & Innovation" of Rennes 1 Foundation.

.

References

Attali,D. (2017) colourpicker: A Colour Picker Tool for Shiny and for Selecting Colours in Plots.

Auguie,B. (2017) gridExtra: Miscellaneous Functions for 'Grid' Graphics.

Benno Haarman,B.C.M. et al. (2015) Feature-expression heat maps - A new visual method to explore complex associations between two variable sets. J. Biomed. Inform., 53, 156–161.

Chang,W. et al. (2019) shiny: Web Application Framework for R.

Chang,W. and Borges Ribeiro,B. (2018) shinydashboard: Create Dashboards with 'Shiny'.

Choi,H. et al. (2011) SAINT: Probabilistic scoring of affinity purificationg-mass spectrometry data. Nat. Methods, 8, 70–73.

Efremova,M. et al. (2019) CellPhoneDB v2.0: Inferring cell-cell communication from combined expression of multi-subunit receptor-ligand complexes. bioRxiv, 680926.

Habib,N. et al. (2017) Massively parallel single-nucleus RNA-seq with DroNc-seq. Nat. Methods, 14, 955–958.

Henglin,M. et al. (2019) A Single Visualization Technique for Displaying Multiple Metabolite–Phenotype Associations. Metabolites, 9, 128.

Husson,F. et al. (2019) FactoMineR: Multivariate Exploratory Data Analysis and Data Mining.

Innes,B.T. and Bader,G.D. (2019) scClustViz – Single-cell RNAseq cluster assessment and visualization [version 2; peer review: 2 approved]. F1000Research, 7.

Knight,J.D.R. et al. (2017) ProHits-viz: A suite of web tools for visualizing interaction proteomics data. Nat. Methods, 14, 645–646.

Lukassen,S. et al. (2018) Characterization of germ cell differentiation in the male mouse through single-cell RNA sequencing OPEN. Sci. Rep., 8, 6521.

Mellacheruvu,D. et al. (2013) The CRAPome: A contaminant repository for affinity purification-mass spectrometry data. Nat. Methods, 10, 730–736.

Ordovas-Montanes,J. et al. (2018) Allergic inflammatory memory in human respiratory epithelial progenitor cells. Nature, 560, 649–654.

Pedersen,T.L. (2019) ggforce: Accelerating 'ggplot2'.

Perrier,V. et al. (2019) shinyWidgets: Custom Inputs Widgets for Shiny.

Stuart,T. et al. (2019) Comprehensive Integration of Single-Cell Data. Cell, 177, 1888-1902.e21.

de Vries,A. and Ripley,B.D. (2016) ggdendro: Create Dendrograms and Tree Diagrams Using 'ggplot2'.

Wei,T. and Simko,V. (2017) corrplot: Visualization of a Correlation Matrix.

Wickham,H. et al. (2019) ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics.

Wickham,H. (2017) reshape2: Flexibly Reshape Data: A Reboot of the Reshape Package.

Wickham,H. and Pedersen,T.L. (2019) gtable: Arrange 'Grobs' in Tables.

Wilkinson,L. and Friendly,M. (2009) The history of the cluster heat map. Am. Stat., 63, 179–184.

Wu,H. et al. (2018) Comparative Analysis and Refinement of Human PSC-Derived Kidney Organoid Differentiation with Single-Cell Transcriptomics. Cell Stem Cell, 23, 869-881.e8.