Violin SuperPlots: visualizing replicate heterogeneity in large data sets.

AUTHOR(S)

Martin Kenny, Ingmar Schoen

Title: Violin SuperPlots: Visualising replicate heterogeneity in large datasets

Authors: Martin Kenny and Ingmar Schoen

Affiliations:

School of Pharmacy and Biomolecular Sciences, Irish Centre for Vascular Biology (ICVB), Royal College of Surgeons in Ireland (RCSI), Dublin, Ireland.

Address correspondence to: Ingmar Schoen (ingmarschoen@rcsi.ie)

Running Head: Violin SuperPlots

Abbreviations:

Letter to the editor

A recent article in MBoC (Goedhart, 2021) presented a web interface for the creation of 'SuperPlots'. SuperPlots were introduced by Lord and colleagues last year (Lord *et al.*, 2020) to visualise both cell-level variability within replicates as well as the experimental reproducibility between replicates in one single plot. Simple bar charts or boxplots of mean or median values from experimental replicates mask the contribution of underlying cell-to-cell variations in individual experiments, whereas pooling cell-level data across replicates overemphasises statistical differences. The SuperPlot put forward by Lord et al. uses a beeswarm plot to display the cell-level data color-coded according to the individual replicates, and overlays the mean (or median) and error bars (standard deviation or confidence intervals) of each replicate (Figure 1a). The new web interface (Goedhart, 2021) offers an online option for researchers to generate beeswarm SuperPlots, as well as RainCloud plots (Allen *et al.*, 2021), using their own data. We welcome the transparency brought by SuperPlots and would like to introduce an augmentation, the Violin SuperPlot, to further simplify visual inspection of raw data containing large sample sizes.

Beeswarm plots are a direct visualisation of the raw data points which sample an underlying parameter distribution. As the number of data points increases, the individual points become indistinguishable when the outline of the beeswarm plot approaches the shape of the underlying parameter distribution. Moreover, the jittered arrangement of color-coded beeswarms in SuperPlots makes it very difficult to identify differences in the replicates' distributions (Figure 1a). Lacking suitable alternatives, researchers have chosen to show the pooled data distribution using a violin plot, which does not contain information about the individual cell distributions within biological replicates (Chavali *et al.*, 2020; Pagès *et al.*, 2020). We thus propose replacing the beeswarm plot with a modified violin plot. A violin plot is essentially a smoothened histogram rotated by 90°, which provides a density estimation of these data (Hintze and Nelson, 1998). In our Violin SuperPlot (Figure 1b), the normalized density estimates of individual replicates are stacked to show how each replicate (color-coded stripe) contributes to the overall density estimate (outline), allowing rapid inspection of experimental variability. These vertical stripes are then overlaid with markers for the central tendency of each distribution (mean or median) and summary statistics (mean and standard error of the mean). Compared to a lesser-known visual representation, the so-called raincloud plot (Allen *et al.*, 2021; Goedhart,

2021), Violin SuperPlots are more compact and concise, thus allowing for rapid visual comparisons and interpretation.

Violin SuperPlots are especially useful for high throughput single cell datasets from microscopy screenings which contain hundreds of cells per experimental replicate (Pepperkok and Ellenberg, 2006; Jones *et al.*, 2008). Certain cell parameters are not necessarily normally distributed. For example, cell spreading area can show one-sided distributions with a tail in either direction, depending on the proportion of spread versus non-spread cells, which may vary upon drug treatment or due to experimental variability (see Figure 1, here from donor to donor). This can be directly appreciated from the width of the stripes in a Violin SuperPlot (Figure 1b) even for experiments containing more than 3 replicates (Figure 1c), but is less clear from the color-coded points of a beeswarm representation (Figure 1a).

Violin SuperPlots are particularly suited for datasets with >10 data points per replicate and up to ~18 biological replicates (Supplementary Figure S1). For less data points (<10) and no more than 3 replicates, a direct depiction of the raw data by a color-coded beeswarm plot might be considered more appropriate than the smoothened density estimate of a violin plot. For many biological replicates (>18), the shape of the individual stripes of a Violin SuperPlot becomes uninformative. In this limiting case, plotting the replicate means together with their summary statistics on top of a violin plot of the pooled data (Chavali *et al.*, 2020; Pagès *et al.*, 2020) provides a suitable compromise. Violin SuperPlots thus do not replace previous SuperPlot formats (Lord *et al.*, 2020; Goedhart, 2021) but rather complement and extend their scope.

To help cell biologists generate Violin SuperPlots from their own data, we have developed a Python-based command-line application built upon libraries that are routinely used for scientific data processing and visualisation (Harris *et al.*, 2020; Virtanen *et al.*, 2020). The application was designed to be accessible for programmers and non-programmers alike, and allows for effortless customization of the generated plots to suit user preferences. The package and supporting documentation are freely available from the PyPI repository and in the Supplementary Material accompanying this article. A basic implementation for MATLAB is also available as Supplementary Material. The software licence also allows the integration of these Violin SuperPlots into a web interface and other data visualisation programs.

We join Goedhart and Lord et al. in encouraging authors to represent data in ways that help the reader to assess biological variation within individual experiments, between biological replicates,

and between conditions. We hope that researchers will find the Violin SuperPlots intuitive and helpful for this purpose.

Acknowledgements

We would like to thank Jonas Ries for contributing to the implementation of Violin SuperPlots in MATLAB, and the anonymous reviewers for their constructive feedback. This work was supported through funding from RCSI (I.S.).

References

Allen, M, Poggiali, D, Whitaker, K, Marshall, TR, and Kievit, RA (2021). Raincloud plots: a multi-platform tool for robust data visualization. Wellcome Open Res 4, 63.

Chavali, M, Ulloa-Navas, MJ, Pérez-Borredá, P, Garcia-Verdugo, JM, McQuillen, PS, Huang, EJ, and Rowitch, DH (2020). Wnt-Dependent Oligodendroglial-Endothelial Interactions Regulate White Matter Vascularization and Attenuate Injury. Neuron 108, 1130-1145.e5.

Goedhart, J (2021). SuperPlotsOfData – a web app for the transparent display and quantitative comparison of continuous data from different conditions. Mol Biol Cell, mbc.E20-09-0583.

Harris, CR et al. (2020). Array programming with NumPy. Nature 585, 357–362.

Hintze, JL, and Nelson, RD (1998). Violin Plots: A Box Plot-Density Trace Synergism. Am Stat 52, 181–184.

Jones, TR, Kang, I, Wheeler, DB, Lindquist, RA, Papallo, A, Sabatini, DM, Golland, P, and Carpenter, AE (2008). CellProfiler Analyst: data exploration and analysis software for complex image-based screens. BMC Bioinformatics 9, 482.

Lord, SJ, Velle, KB, Mullins, RD, and Fritz-Laylin, LK (2020). SuperPlots: Communicating reproducibility and variability in cell biology. J Cell Biol 219.

Pagès, D-L et al. (2020). Cell clusters adopt a collective amoeboid mode of migration in confined non-adhesive environments. BioRxiv, 1–17.

Pepperkok, R, and Ellenberg, J (2006). High-throughput fluorescence microscopy for systems biology. Nat Rev Mol Cell Biol 7, 690–696.

Virtanen, P et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17, 261–272.

Figures

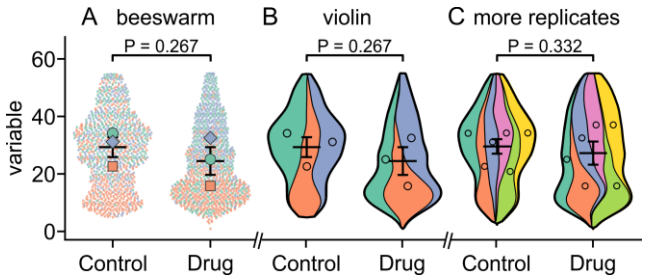

Figure 1. Violin SuperPlots for the visualisation of replicate heterogeneity in large datasets. (A) Beeswarm SuperPlots show cell-level (technical replicates) data colour-coded by experimental (biological) replicate. Distributions of individual replicates can be difficult to interpret due to the density and jitter of the data points. The plot was created using the SuperPlotOfData web app. (B) Violin SuperPlots depict cell-level data from each replicate as stripes in a compound violin plot. Same data as in A. (C) The number of replicates (in this case 6) in Violin SuperPlots can be increased without compromising readability. Symbols: means of experimental replicates. Lines: mean and standard error of the mean of the replicate means. Statistical test: paired student's t-test. Data shown: spreading area (μm^2) of human platelets seeded on fibrinogen-coated coverslips for 60 minutes in the presence/absence of 40 μM blebbistatin.