Violin SuperPlots: visualizing replicate heterogeneity in large data sets

Martin Kenny and Ingmar Schoen*
School of Pharmacy and Biomolecular Sciences, Irish Centre for Vascular Biology, Royal College of Surgeons in Ireland, Dublin 2, Ireland

To the editor:

A recent article in *Molecular Biology of the Cell* (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 visualize 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 overemphasizes 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 (SD 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 visualization of the raw data points that sample an underlying parameter distribution. As the number of data points increases, the individual points become indistinguishable while 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 that 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° that 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 SEM). 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 data sets from microscopy screenings that 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 nonspread 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 three replicates (Figure 1C), but is less clear from the color-coded points of a beeswarm representation (Figure 1A).

Violin SuperPlots are particularly suited for data sets with >10 data points per replicate and up to ~18 biological replicates (Supplemental Figure S1). For fewer data points (<10) and no more than three 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 visualization (Harris *et al.*, 2020; Virtanen *et al.*, 2020). The

DOI:10.1091/mbc.E21-03-0130
*Address correspondence to: Ingmar Schoen (ingmarschoen@rcsi.ie).

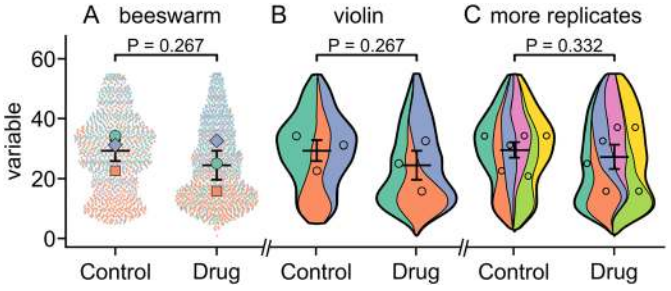

FIGURE 1: Violin SuperPlots for the visualization of replicate heterogeneity in large data sets. (A) Beeswarm SuperPlots show cell-level (technical replicates) data color-coded by experimental (biological) replicate. Distributions of individual replicates can be difficult to interpret due to the density and jitter of the data points. This 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 six) in Violin SuperPlots can be increased without compromising readability. Symbols: means of experimental replicates. Lines: mean and SEM 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 min in the presence/absence of 40 μM blebbistatin.

application was designed to be accessible for programmers and nonprogrammers alike and allows for effortless customization of the generated plots to suit user preferences (Supplemental Figures S2–S4). The package and supporting documentation are freely available from the PyPI repository and in the Supplemental Material accompanying this Letter. A basic implementation for MATLAB is also available as Supplemental Material. The software license also allows the integration of these Violin SuperPlots into a web interface and other data visualization programs.

We join Goedhart (2021) and Lord *et al.* (2020) 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.

ACKNOWLEDGMENTS

We 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 the Royal College of Surgeons in Ireland (I.S.).

REFERENCES

Allen M, Poggiali D, Whitaker K, Marshall TR, van Langen J, 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, 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 32, 470–474.

Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, *et al.* (2020). Array programming with NumPy. Nature 585, 357–362.

Hintze JL, 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, 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, Fritz-Laylin LK (2020). SuperPlots: communicating reproducibility and variability in cell biology. J Cell Biol 219, e202001064.

Pagès D-L, Dornier E, De Seze J, Wang L, Luan R, Cartry J, Canet-Jourdan C, Raingeaud J, Voituriez R, Coppey M, *et al.* (2020). Cell clusters adopt a collective amoeboid mode of migration in confined non-adhesive environments. BioRxiv, doi: https://doi.org/10.1101/2020.05.28.106203.

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

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, *et al.* (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17, 261–272.