Variation in ancestry proportions shown through violin plots

Understanding variation in biology is something that a majority of biologists (human or otherwise) dedicate their careers to. Visually displaying this variation is a key factor in this understanding. If we want to know how variation exists among populations, there are a few ways that we can look at this. One of the main ways to look at variation is through boxplots. Boxplots are also known as box and whisker plots, and they give valuable information about the data. Boxplots have 5 main components – the minimum value, the first quartile, the median, the third quartile, and the maximum. Through these components, one can begin to develop an understanding of the data and what sort of variation exists in the populations of interest. In addition to boxplots, violin plots can be of interest to biologists as well. A violin plot can show the density of the data at a certain value as well as showing outliers in the data. Combined, these two plot types can be extremely valuable to giving a visual understanding of the data of interest.

In my specific tutorial, I am using four different data sources for two sets that represent populations of interest. I am aiming to show the variation of a certain ancestry in 7 populations that are from the same country. The first pair of data is an admixture/ancestry proportions file and an ID file for Population 1 and reference populations. The data consists of 1067 observations for one variable (ID) and 1067 observations for five variables (ancestries). The second pair of data is, again, ancestry proportions and ID files for Populations 2-7 and reference populations. The reason that there are two pairs of data are due to sampling locations and sampling dates. Different populations were collected during different field seasons, but to analyze our data properly we needed to combine these two pairs of data for the overall violin plot.

The goal of the tutorial will be to wrangle the 4 different data types into one file focusing on only the 7 populations of interest and only the ancestry of interest. I will do this by filtering out the “Reference” populations in the ID files, and filtering out any ancestry column that isn’t “KHS”. I will use ggplot for the boxplots and violin plots and the dplyr package for data manipulation and wrangling. The final output will be a violin plot with boxplots layered on top for a full visual summary of the data.

Hintze, J. L., & Nelson, R. D. (1998). Violin Plots: A Box Plot-Density Trace Synergism. *The American Statistician*, *52*(2), 181–184. https://doi.org/10.1080/00031305.1998.10480559

Kenny, M., & Schoen, I. (2021). Violin SuperPlots: Visualizing replicate heterogeneity in large data sets. *Molecular Biology of the Cell*, *32*(15), 1333–1334. https://doi.org/10.1091/mbc.E21-03-0130

Lord, S. J., Velle, K. B., Mullins, R. D., & Fritz-Laylin, L. K. (2020). SuperPlots: Communicating reproducibility and variability in cell biology. *Journal of Cell Biology*, *219*(6), e202001064. https://doi.org/10.1083/jcb.202001064

Molina, E., Viale, L., & Vázquez, P. (2022). How should we design violin plots? *2022 IEEE 4th Workshop on Visualization Guidelines in Research, Design, and Education (VisGuides)*, 1–7. https://doi.org/10.1109/VisGuides57787.2022.00006

Tanious, R., & Manolov, R. (2022). Violin Plots as Visual Tools in the Meta-Analysis of Single-Case Experimental Designs. *Methodology*, *18*(3), Article 3. https://doi.org/10.5964/meth.9209

Weissgerber, T. L., Winham, S. J., Heinzen, E. P., Milin-Lazovic, J. S., Garcia-Valencia, O., Bukumiric, Z., Savic, M. D., Garovic, V. D., & Milic, N. M. (2019). Reveal, Don’t Conceal. *Circulation*, *140*(18), 1506–1518. https://doi.org/10.1161/CIRCULATIONAHA.118.037777