POINTS OF VIEW

Avoiding color

Last month I wrote about color blindness and ways to make information accessible to individuals with color vision deficiencies. I would like to continue by considering graphical alternatives to color that could improve the overall clarity and utility of data displays.

The primary use of color in research is to convey information. When used effectively, color can simplify a complex analysis task. When misused, it can bias a reader's perception of the underlying data. For example, when color gradients indicating relative quantity contain abrupt transitions, specific numerical ranges can be preferentially accentuated (**Fig. 1a**). Edward Tufte advises us that color used poorly is worse than no color at all; his motto is: "Above all, do no harm"[1]. Color can cause the wrong information to stand out and make meaningful information difficult to see. Furthermore, the overuse of color can produce visual clutter akin to signage in Times Square or Piccadilly Circus with countless elements competing for our attention.

In addition to limiting accessibility, there are several other disadvantages to using color to present data. I showed how the visual phenomenon resulting from the interaction of color can cause the same color in heatmaps to appear different[2]. Color is a relative medium. When we pair hues varying greatly in saturation or value (lightness), we can unintentionally produce presentations that are lopsided. In **Figure 1b**, the light blue bands appear under-represented partially because they are lighter than the other colors as evident by looking at the key in grayscale (**Fig. 1b**). Color can also elicit size biases; some people find equal areas filled with vibrant colors seem to be more dissimilar than when less saturated colors are used.

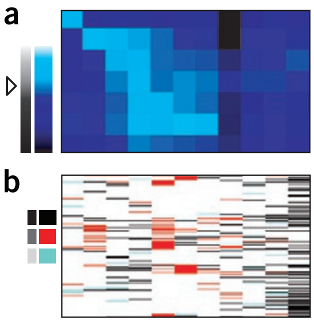

Figure 1 | Color can mask data. (**a**) Color scale with sharp transition in hue and value (arrow) can exaggerate specific data ranges. (**b**) Juxtaposing colors highly varying in saturation and value can make aspects of the data appear under-represented (light blue).

Although color is an attractive choice for conveying information, it may not be the best visual cue to bring out relevant trends. Color hue can be such a potent differentiator that using size, shape, texture, length, width, orientation, curvature and intensity to encode information may enable more aspects of the data to be discriminable. Our choice of graphical cues should depend on what we and others need to see to reliably pick out patterns.

In one project at the Broad Institute, researchers wanted to understand the evolution of molecular networks by studying gene expression in yeast. They had time course data for about a dozen species. The researchers were interested in comparing expression profiles across genes and species. With their data displayed as heatmaps, it is difficult to characterize the differences between profiles (**Fig. 2a**). Redrawing the data as line graphs and shading the area under the curve better support the visual task of comparing patterns for mirror symmetry and peak shift (**Fig. 2b**). To gauge conservation across metabolic pathways, the researchers calculate a correlation score accounting for all species for each node in the network and assign color to score (**Fig. 2c**). As it is difficult to sequence color hues, mapping the data to length and position makes it easier to see points of high and low correlation (**Fig. 2d**). The compact format allowed data for both metabolites and genes to be displayed (**Fig. 2d**). The visual complexity that comes from too many colors makes it difficult to also show the metabolite data in the original scheme (**Fig. 2c**).

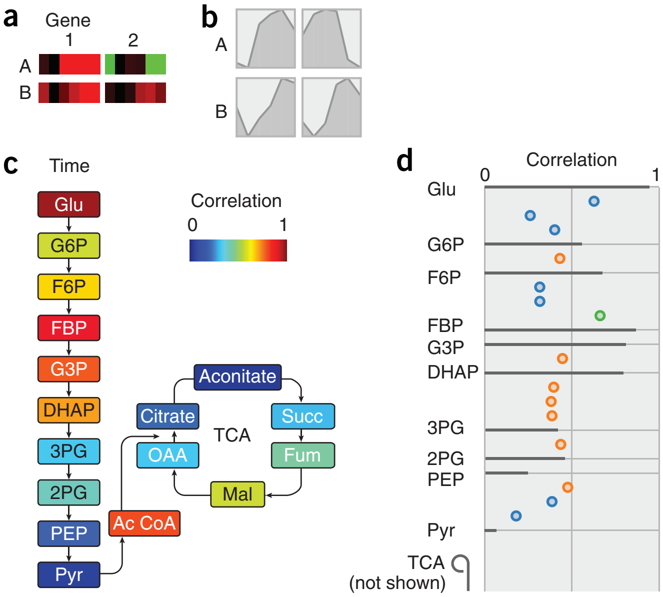

Figure 2 | Color can limit accessibility and hinder analysis. (**a**) Heatmap representation of time series data for species A and B. (**b**) Filled line charts of data from **a** facilitate profile comparison. (**c**) Color hue indicates correlation score for metabolites in glycolysis (boxes). Enzymes are shown as arrows. (**d**) Replacing color encoding from **c** with bar length for metabolites and position of circles on the *x* axis for enzymes increases data density and makes rank ordering easy. Color indicates directionality of enzymatic activity. Visualization technique is from reference 3.

Color is often our first choice when it comes to showing data. Depending on the fundamental visual task required for analysis, basic diagrammatic marks may do a better job of revealing data structures. I have seen squiggly lines used effectively to denote several data dimensions at once. Although color is inextricably tied to what many of us consider to have high visual impact, expressiveness relies primarily on one's graphical selection, whereas effectiveness also depends on the capabilities of the perceiver.

Bang Wong

1. Tufte, E. *Envisioning Information* (Graphics Press, Cheshire, Connecticut, USA, 1990).
2. Wong, B. *Nat. Methods* **7**, 665 (2010).
3. Meyer, M. *et al. Proc. EuroVis* **29**, 1043–1052 (2010).

Bang Wong is the creative director of the Broad Institute of the Massachusetts Institute of Technology & Harvard and an adjunct assistant professor in the Department of Art as Applied to Medicine at The Johns Hopkins University School of Medicine.