POINTS OF VIEW

Mapping quantitative data to color

Data structure informs choice of color maps.

Data can be classified in many ways. One useful method of classifying data for visualization is to distinguish between those with and without an inherent order. For example, a set of species (such as *Escherichia coli*, *Drosophila melanogaster* and *Homo sapiens*) has no intuitive ordering and is considered 'categorical data', whereas a list of gene expression values is 'ordered data' because we can sort them from lowest to highest. In a previous column, we described methods for color-coding categorical data (August 2010)[1]. Here we focus on creating color maps for quantitative data.

Color is arguably one of the most important graphical assets for data presentation, from medical imaging to pie charts. By varying just three primary components of color (hue, saturation and lightness), color can fulfill a number of fundamental communication needs: to label, to show quantities, to represent or simulate reality, and to enliven or decorate. It is imperative that we choose color purposefully to highlight the salient features of the data we intend to depict. Even though color encoding does not result in the most accurate visual representation of quantitative data, color is often the best choice for compact visualizations of large data sets.

Unlike categorical data, the elements of quantitative data can be placed on a numerical scale that describes their relative position and size with respect to one another. This interrelationship of quantitative data requires that we exercise care in designing color maps that are perceptually consistent with the range and change in magnitude found in the data.

When depicting quantitative data, it is useful to first define the key regions or points in the data range that we intend to highlight before designing a color-coding scheme that varies the three components of color. Often this requires determining the aspects of the data we want to make apparent. In many cases, the meaningful range will be the extremes—the minimum and/or maximum values. Additionally, there can be numerical values between the extremes with special meaning, such as zero. In some cases, this number could be unique to the defined data range, such as 'sea level' for maps or 32° on the Fahrenheit temperature scale.

Although color hue is well suited for categorical data, it tends to be impractical for quantitative data. With quantitative data, we principally rely on color value and reserve hue to indicate different segments of the data range. When plotting data with only positive or negative values, an intuitive encoding is a sequential color map that varies only the lightness from 10% to 90% black (**Fig. 1a**). Such a color progression produces even transitions throughout the range. There are two possible options for fitting such a color map to the data: we can translate the ends of the color gradient to (i) zero and the theoretical maximum value or (ii) the observed minimum and maximum. The former approach allows us to interpret the data in the context of the theoretical data range (**Fig. 1a**). However, if higher contrast is needed from the graphical representation and zero is irrelevant as a reference point, then it is reasonable to map the lowest observed value to the lightest color and the highest observed value to the darkest color (**Fig. 1b**).

In circumstances where the data have more than two regions of interest, it is necessary to design a color schema with multiple facets. A common scenario involves data containing both positive and negative values, in which the lower and upper ends of the distribution as well as zero need to be distinguished. In this case, a diverging (or bipolar) color schema that employs both color hue and color saturation is effective. Use color hue to make a distinction between positive and negative values (for example, red and blue) and color saturation to indicate the relative scale, with more saturated color depicting values of greater magnitude and no saturation representing zero (**Fig. 1c**).

The interpretation of zero or other key values can further influence the choice of color keys. Geographical elevation maps use keys that make the zero crossing visually explicit (**Fig. 1d**). This is achieved by using different colors for the values immediately below and above zero, respectively. Whether such a color map is appropriate for the data depends on how variable the data are and the meaning of zero for the interpretation.

It is essential to select the type of the color maps appropriate for the data. Some analytical software tools use a divergent color map as a default. When this is inadvertently applied to data ranges without a zero crossing, the data may be misrepresented because an increase in data values might not be reflected by an increase in color saturation (**Fig. 1e**). When designing color maps, there are two resources we like that do not require the user to supply all of the color expertise. They are the Pennsylvania State University's ColorBrewer (http://colorbrewer2.org/) and NASA's Color Tool (http://colorusage.arc.nasa.gov/ColorTool.php#1).

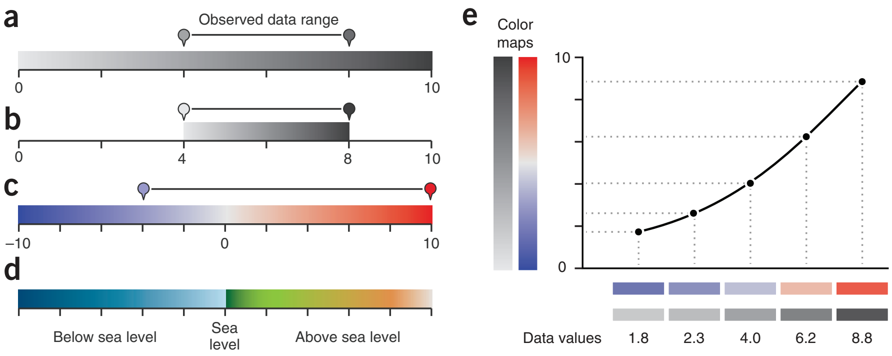

Figure 1 | Color maps. (**a**) Sequential color gradient from 10% to 90% black. (**b**) A sequential color schema mapped to observed data range. (**c**) Divergent color gradient varying in hue and saturation. (**d**) Blended-hue color map. (**e**) Schematic illustration of a misleading representation due to misaligned data and color properties.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Nils Gehlenborg & Bang Wong

Nils Gehlenborg is a research associate at Harvard Medical School and the Broad Institute. Bang Wong is the creative director of the Broad Institute and an adjunct assistant professor in the Department of Art as Applied to Medicine at The Johns Hopkins University School of Medicine.

1. Wong, B. *Nat. Methods* **7**, 573 (2010).